Bayesian Inverse Problems

Lecture notes, Lent term 2019
University of Cambridge

Hanne Kekkonen

November 6, 2019

Contents

Abstract

Inverse problems arise from the need to gain information about an unknown object of interest from given indirect measurements. Inverse problems have several applications varying from medical imaging and industrial process monitoring to ozone layer tomography and modelling of financial markets. The common feature for inverse problems is the need to understand indirect measurements and to overcome extreme sensitivity to noise and modelling inaccuracies. In this course we employ probabilistic approach to inverse problems to find stable and meaningful solutions that allow us quantify how inaccuracies in the data or model affect the obtained estimate.

1 Bayesian approach to discrete inverse problems

1.1 Introduction

We start by considering the problem of finding $u \in \mathbb{R}^d$ that satisfies the equation

$$m_0 = Au, \tag{1.1}$$

where $m_0 \in \mathbb{R}^k$ is given. We refer to m_0 as observed data or measurement and u as an unknown. The physical phenomena that relates the unknown and the measurement is modelled by a matrix $A \in \mathbb{R}^{k \times d}$. In real life the perfect data given in (1.1) is perturbed by noise and we observe measurements

$$m = Au + n, \tag{1.2}$$

where $n \in \mathbb{R}^k$ represents the observational noise.

We are interested in ill-posed inverse problems, where the inverse problem is more difficult to solve than the direct problem of finding m when u is given. To explain this we first need to introduce well-posedness as defined by Jacques Hadamard:

I Existence: There exists at least one solution.

II Uniqueness: There is at most one solution.

III Stability: The solution depends continuously on data.

The direct or forward problem is assumed to be well-posed. The inverse problems are generally ill-posed and break at least one of the above conditions.

1. Assume that $d < k$ and $A : \mathbb{R}^d \to \mathcal{R}(A) \subsetneq \mathbb{R}^k$, where the range of A is a proper subset of $\mathbb{R}^k$. Furthermore, we assume that A has a unique inverse $A^{-1} : \mathcal{R}(A) \to \mathbb{R}^k$. Because of the noise in the measurement $m \notin \mathcal{R}(A)$ so that simply inverting A with the data given in (1.2) is not possible. Note that usually only the statistical properties of the noise n are known so we cannot just subtract it.

2. Assume next that $d > k$ and $A : \mathbb{R}^d \to \mathbb{R}^k$, in which case the system is underdetermined. We then have more unknowns than equations which means that there are several possible solutions.

3. Consider next case $d = k$ and there exist $A^{-1} : \mathbb{R}^k \to \mathbb{R}^k$ but the condition number $\kappa = \lambda_1/\lambda_k$, where λ_1 and λ_k are the biggest and smallest eigenvalues of A, is very large. Such a matrix is said to be ill-conditioned and is almost singular. In this case the problem is sensitive even to smallest errors in the measurement. Hence the naive reconstruction $\widetilde{u} = A^{-1}f_n = u + A^{-1}n$ does not produce a meaningful solution but will be dominated by $A^{-1}n$. Note that $\|A^{-1}n\|_2 \approx \|n\|_2/\lambda_k$ can be arbitrarily large.

The last part illustrates one of the key perspectives of inverse problem theory; How can we stabilise the reconstruction process while maintaining acceptable accuracy?

Next we will take a look at some examples of inverse problems to see what kind of challenges we face when trying to solve them.

Example 1.1. The deblurring (or deconvolution) problem of recovering an input signal u form an observed signal

$$m(t) = \int_{-\infty}^{\infty} a(t-s)u(s)ds + n(t)$$

occurs in many imaging, and image- and signal processing applications. Here the function a is known as the blurring kernel.

The noiseless data is given by $m_0(t) = \int_{-\infty}^{\infty} a(t-s)u(s)ds$ and its Fourier transform is $\widehat{m}_0(\xi) = \int_{-\infty}^{\infty} e^{-i\xi t}m_0(t)dt$. The convolution theorem implies

$$\widehat{m}_0(\xi) = \widehat{a}(\xi)\widehat{u}(\xi),$$

and hence by inverse Fourier transform

$$u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\xi}\frac{\widehat{m}_0(\xi)}{\widehat{a}(\xi)}d\xi.$$

However, we can only observe noisy measurements and hence we have on the frequency domain $\widehat{m}(\xi) = \widehat{a}(\xi)\widehat{u}(\xi) + \widehat{n}(\xi)$. The estimate u_{est} based on the convolution theorem is given by

$$u_{est}(t) = u(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\xi}\frac{\widehat{n}(\xi)}{\widehat{a}(\xi)}d\xi,$$

which is often not even well defined, since usually the kernel a decreases exponentially (or has compact support), making the denominator small, whereas the Fourier transform of the noise will be non-zero.

Example 1.2. Next we study the problem of recovering the initial condition u of the heat equation from a noisy observation m of the solution at some time $T > 0$. We consider the heat equation on a torus $\mathbb{T}^d$, with Dirichlet boundary conditions

$$\begin{cases} \frac{dv}{dt} - \Delta v = 0 & \text{on } \mathbb{T}^d \times \mathbb{R}_+ \\ v(x,t) = 0 & \text{on } \partial\mathbb{T}^d \times \mathbb{R}_+ \\ v(x,T) = m_0(x) & \text{on } \mathbb{T}^d \\ v(x,0) = u(x) & \text{on } \mathbb{T}^d \end{cases}$$

where Δ denotes the Laplace operator and $\mathcal{D}(\Delta) = H_0^1(\mathbb{T}^d) \cap H^2(\mathbb{T}^d)$. Note that the operator $-\Delta$ is positive and self-adjoint on Hilbert space $\mathcal{H} = L^2(\mathbb{T}^d)$.

Given a function $u \in L^2(\mathbb{T}^d)$ we can decompose it as a Fourier series

$$u(x) = \sum_{n\in\mathbb{Z}^d} u_n e^{2\pi i\langle n,x\rangle},$$

where $u_n = \langle u, e^{2\pi i\langle n,x\rangle}\rangle$ are the Fourier coefficients, and the identity holds for almost every $x \in \mathbb{T}^d$. The L^2 norm of u is given by the Parseval's identity $\|u\|_{L^2}^2 = \sum u_n$. Remember that the Sobolev space $H^s(\mathbb{T}^d)$, $s \in \mathbb{N}$, consist of all $L^2(\mathbb{T}^d)$ integrable functions whose α^{th} order weak derivatives exist and are $L^2(\mathbb{T}^d)$ integrable for all $|\alpha| \leq s$. The fractional Sobolev space $H^s(\mathbb{T}^d)$ is given by the subspace of functions $u \in L^2(\mathbb{T}^d)$, such that

$$\|u\|_{H^s}^2 = \sum_{n\in\mathbb{Z}^d} (1 + 4\pi^2|n|^2)^s |u_n|^2 < \infty. \tag{1.3}$$

Note that for a positive integer s, the above definition agrees with the definition given using the weak derivatives. For $s < 0$, we define $H^s(\mathbb{T}^d)$ via duality or as the closure of $L^2(\mathbb{T}^d)$ under the norm (1.3). The resulting spaces are separable for all $s \in \mathbb{R}$.

The eigenvectors of $-\Delta$ in $\mathbb{T}^d$ form the orthonormal basis of $L^2(\mathbb{T}^d)$ and the eigenvalues are given by $4\pi^2|n|^2$, $n \in \mathbb{Z}^d$. We can also work on real-valued functions where the eigenfunctions $\{\phi_j\}_{j=1}^\infty$ comprise sine and cosine functions. The eigenvalues of $-\Delta$, when ordered on a one-dimensional lattice, then satisfy $\lambda_j \asymp j^{\frac{2}{d}}$. The notation $\asymp$ means that there exist constants $C_1, C_2 > 0$, such that $C_1 j^{\frac{2}{d}} \leq \lambda_j \leq C_2 j^{\frac{2}{d}}$.

The solution to the forward heat equation can be written as

$$v(t) = \sum_{j=1}^\infty u_j e^{-\lambda_j t}\phi_j.$$

We notice that

$$\|v(t)\|_{H^s}^2 \asymp \sum_{j=1}^\infty j^{\frac{2s}{d}} e^{-2\lambda_j t}|u_j|^2 = t^{-s}\sum_{j=1}^\infty (\lambda_j t)^s e^{-2\lambda_j t}|u_j|^2 \leq Ct^{-s}\sum_{j=1}^\infty |u_j|^2 = Ct^{-s}\|u\|_{L^2}$$

which implies that $v(t) \in H^s(\mathbb{T}^d)$ for all $s > 0$.

We now have observation model

$$m = Au + n,$$

where $A = e^{\Delta}$ and n is the observational noise. The noise is not usually smooth (the often assumed white noise is not even an L^2 function) and hence measurement m is not in the image space $\mathcal{D}(e^{\Delta}) \subset \cap_{s>0} H^s(\mathbb{T}^d)$.

A deterministic way of achieving a unique and stable solution for the problem (1.2) is to use regularisation theory. In the classical Tikhonov regularisation a solution is attained by solving

$$\min_{u \in \mathbb{R}^d} \left(\|Au - m\|^2 + \alpha \|Lu\|^2 \right). \tag{1.4}$$

The solution to the above is given by

$$u_n^\alpha = (A^\top A + \alpha L^\top L)^{-1} A^\top m. \tag{1.5}$$

Here α acts as a tuning parameter balancing the effect of the data fidelity term $\|Au - m\|_2^2$ and the stabilising regularisation term $\|u\|_2^2$. Regularisation theory is discussed in more detail in the course *Inverse Problems in Imaging.*

In this course we concentrate on Bayesian inversion. The idea of statistical inversion methods is to rephrase the inverse problem as a question of statistical inference. We consider problem

$$m = Au + \eta, \tag{1.6}$$

where the measurement, unknown and noise are now modelled as random variables. Let $\Omega = \Omega_1 \times \Omega_2$ be our probability space. Then $u : \Omega_1 \to \mathbb{R}^d$, $\eta : \Omega_2 \to \mathbb{R}^k$ and $m : \Omega \to \mathbb{R}^k$.

This approach allows us to model the noise through its statistical properties. We can also encode our *a priori* knowledge of the unknown in form of a probability distribution that assigns higher probability to those values of u we expect to see. Note that the above mentioned regularisation method produces a single estimate of the unknown while the solution to (1.6) is so-called *posterior distribution*, which is the conditional probability distribution of u given a measurement m. This distribution can then be used to obtain estimates that are most likely in some sense. The great advance of the method is, however, that it automatically delivers a quantification of uncertainty, obtained by assessing the spread of the posterior distribution.

We recall the Bayes' formula that states

$$\mathbb{P}(u \in A \,|\, m \in B) = \frac{\mathbb{P}(m \in B \,|\, u \in A)\mathbb{P}(u \in A)}{\mathbb{P}(m \in B)},$$

where A and B are some measurable sets. We would like to solve an inverse problem "approximate u when a measurement $m = m(\omega)$ is given", that is, we would like to condition $u \in A$ with a single realisation of m. To do this we need to we start with some modern probability theory.

1.2 A brief introduction to probability theory

A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ the σ-algebra of events and $\mathbb{P}$ the probability measure. A measure is called σ-finite if Ω is a countable union of measurable sets with finite measure. Lebesgue measure on $\mathbb{R}^d$ is an example of a σ-finite measure. One intuitive way of thinking σ-algebras in probability theory is that they describe information. The σ-algebra contains the subsets representing the events for which we can decide, after the observation, whether they happened or not. Hence $\mathcal{F}$ represents all the information we can get from an experiment in $(\Omega, \mathcal{F}, \mathbb{P})$ while a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ represents partial information.

Let $(X, \mathcal{B}(X))$ be a measurable space, with $\mathcal{B}(X)$ denoting the Borel σ-algebra generated by the open sets. We call a measurable mapping $x : \Omega \to X$ a random variable. The random variable x induces the following probability measure on X

$$\mu(A) = \mathbb{P}(x^{-1}(A)) = \mathbb{P}(\omega \in \Omega : x(\omega) \in A), \quad A \in \mathcal{B}(X).$$

The measure μ is called the probability distribution of x and we will denote $x \sim \mu$.

Let μ and ν be two measures on the same measure space. Then μ is absolutely continuous with respect to (dominated by) ν if $\nu(A) = 0$ implies that $\mu(A) = 0$. We denote this by $\mu \ll \nu$. Measures μ and ν are said to be equivalent if $\mu \ll \nu$ and $\nu \ll \mu$. If μ and ν are supported on disjoint sets they are called mutually singular.

Theorem 1.3 (Radon-Nikodym Theorem)**.** *Let μ and ν be two measures on the same measure space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$ and ν is σ-finite then there exists a unique function $f \in L^1_\nu$ such that for any measurable set $A \in \mathcal{F}$,*

$$\mu(A) = \int_A f d\nu.$$

The unique $f \in L^1_\nu$ in the above theorem is called the Radon-Nikodym derivative of μ with respect to ν and is denoted by $\frac{d\mu}{d\nu}$. The following example shows how Radon-Nikodym Theorem can be used to define probability density for a measure on a finite space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Example 1.4. Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\mu \ll \nu_L$, where ν_L is the standard Lebesgue measure on $\mathbb{R}^d$. Since ν_L is σ-finite we can use Theorem 1.3 and conclude that there exists such $f \in L^1(\mathbb{R}^d)$ that, for any $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mu(A) = \int_A f(t) dt.$$

The function f is called the probability density of $x \sim \mu$.

The σ-algebras we use are often generated by random variables. If $x : \Omega \to X$ then $\sigma(x)$ denotes the smallest σ-algebra containing preimages $x^{-1}(A)$ of measurable sets $A \in \mathcal{B}(X)$. Observing the value of x corresponds of knowing, with every $A \in \mathcal{B}(X)$, whether $x(\omega) \in A$. Note that $\sigma(x) \subset \mathcal{F}$ where, according to the information interpretation, $\mathcal{F}$ represents "full information" (all events on our probability space).

Definition 1.5. *Let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. We call a $\mathcal{G}$-measurable function $y : \Omega \to X$ a conditional expectation of $x : \Omega \to X$ with respect to $\mathcal{G}$ if*

$$\int_G x d\mathbb{P} = \int_G y d\mathbb{P}$$

for all $G \in \mathcal{G}$ and write $\mathbb{E}(x \,|\, \mathcal{G}) = y$.

Note that the measurability with respect to $\mathcal{G}$ is a stronger assumption than measurability with respect to $\mathcal{F}$ since there are fewer choices for the preimages of y. Even though the definition of $\mathbb{E}(x \,|\, \mathcal{G})$ resembles that of $\mathbb{E}(x \,|\, G)$ for an event G these are very different objects. The first is a $\mathcal{G}$-measurable function $\Omega \to X$ while the second is an element in X.

We can also consider conditional expectation of the form $\mathbb{E}(f(x) \,|\, \mathcal{G})$ which leads us to conditional probability.

Definition 1.6. *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional probability for $A \in \mathcal{B}(X)$, given $\mathcal{G}$ is defined by*

$$\mathbb{P}(A \,|\, \mathcal{G}) = \mathbb{E}(\mathbb{1}_A \,|\, \mathcal{G}).$$

It is tempting to try to interpret the map $A \to \mathbb{P}(A \,|\, \mathcal{G})(\omega)$ as a probability measure for a fixed $\omega \in \Omega$. However $\mathbb{P}(A \,|\, \mathcal{G})$ is defined only up to $\mathbb{P}$ almost everywhere.

Definition 1.7. *A family of probability distributions $(\mu(\cdot, \omega))_{\omega \in \Omega}$ on $(X, \mathcal{B}(X))$ is called a regular conditional distribution of x, given $\mathcal{G} \subset \mathcal{F}$, if*

$$\mu(A, \cdot) = \mathbb{E}(\mathbb{1}_A(x) \,|\, \mathcal{G}) \ a.s$$

for every $A \in \mathcal{B}(X)$.

Theorem 1.8. *Let $x : \Omega \to X$ be a random variable and $\mathcal{G} \in \mathcal{F}$ a sub-σ-algebra. Then there exists a regular conditional distribution $(\mu(\cdot, \omega))_{\omega \in \Omega}$ of x given $\mathcal{G}$.*

Let $\sigma(m) \subset \mathcal{F}$ be the σ-algebra generated by a random measurement m. We can then use the regular conditional probability measure

$$\pi_{post}(A, m(\omega)) = \mathbb{E}(\mathbb{1}_A(u) \,|\, \sigma(m))(\omega)$$

as a posterior measure and identify this with $\pi_{post}(A, m) = \pi_{prior}(A \,|\, m)$.

For further information see e.g.For further information see e.g. [8].

1.3 Bayes' formula

We can now return to the problem of "approximate u given a measurement $m = Au + \eta$" using a posterior distribution that is a regular conditional distribution. We assume that u follows a prior Π with Lebesgue density $\pi(u)$. The noise η is assumed to be independent of u and distributed according to P_0 with Lebesgue density $\rho(\eta)$. Then $m \,|\, u$ can be found by simply shifting P_0 by Au to measure P_u, which has Lebesgue density $\rho_u(m) = \rho(m - Au)$. It follows that $(u, m) \in \mathbb{R}^d \times \mathbb{R}^k$ is a random variable with Lebesgue density $\nu(u, m) = \rho(m - Au)\pi(u)$.

Theorem 1.9 (Bayes' Theorem)**.** *Assume that*

$$Z(m) = \int_{\mathbb{R}^d} \rho(m - Au)\pi(u)du > 0.$$

Then $u \,|\, m$ is a random variable with Lebesgue density $\pi^m(u)$ given by

$$\pi^m(u) = \pi(u \,|\, m) = \frac{1}{Z(m)}\rho(m - Au)\pi(u).$$

Let us take a closer look to what the above theorem means in our inverse problems settings.

i) $\pi(u)$ is called *prior density*. The prior should be independent of the measurement and assign higher probability to those values of u we expect to see.

ii) $\rho(m - Au)$ is the *likelihood* which measures the data misfit.

iii) $\pi^m(u)$ is called *posterior density* and it gives a solution to the inverse problem (1.6) by updating the prior with a given measurement.

iv) $Z(m)$ is the probability of m and plays the role of normalising constant.

v) We define

$$\Phi(u; m) = -\log \rho(m - Au)$$

and call Φ *potential*.

vi) Let Π^m and Π be measures on $\mathbb{R}^d$ with densities π^m and π respectively. Then Theorem 1.9 can be rewritten as

$$\frac{d\Pi^m}{d\Pi}(u) = \frac{1}{Z(m)}\exp(-\Phi(u; m)),$$
$$Z(m) = \int_{\mathbb{R}^d} \exp(-\Phi(u; m))d\Pi(u).$$

Note that this means the posterior is absolutely continuous with respect to the prior and the Radon-Nikodym derivative is proportional to the likelihood.

When stated as in vi) the formula has a natural generalisation to infinite dimensions where there are no densities ρ and π with respect to Lebesgue measure but where Π^m has a Radon-Nikodym derivative with respect to Π.

Remark 1.10. In Example 1.4 we defined density f of a measure μ in $\mathbb{R}^d$, which is absolutely continuous with respect to Lebesgue measure ν_L. Strictly speaking $f(x) = \frac{d\mu}{d\nu_L}(x)$ is a probability density function with respect to Lebesgue measure.

It is also possible to find the density of μ with respect to a Gaussian measure. Let $\mu_0 \sim \mathcal{N}(0, I)$ denote the standard Gaussian measure in $\mathbb{R}^d$. Then

$$\mu_0(dx) = \frac{1}{(2\pi)^{d/2}}\exp\bigg(-\frac{1}{2}|x|^2\bigg)dx.$$

Thus the density of μ with respect to μ_0 is

$$f_G(x) = (2\pi)^{d/2} \exp\left(\frac{1}{2}|x|^2\right) f(x).$$

We then have identities

$$\mu(A) = \int_A f_G(x)\mu_0(dx) \quad \text{and} \quad \frac{d\mu}{d\mu_0}(x) = f_G(x). \tag{1.7}$$

Note that infinite-dimensional Gaussian measure is well-defined (we return to this in Section 2.4) while there is no infinite dimensional Lebesgue measure. Many measures have a Radon–Nikodym derivative with respect to an infinite-dimensional Gaussian measure and hence formulation (1.7) can be generalised to infinite-dimensional settings while the Lebesgue density can not.

Example 1.11. We start by studying the case $u \in \mathbb{R}$ and $m \in \mathbb{R}^k$, $k \geq 1$. The measurement is defined by

$$m = Au + \eta,$$

where $A \in \mathbb{R}^k \setminus \{0\}$ and $\eta \sim \mathcal{N}(0, \delta^2 I)$. We model the unknown u by a Gaussian measure $\mathcal{N}(0,1)$. Then

$$\pi^m(u) \propto \exp\left(- \frac{1}{2\delta^2}\|m - Au\|^2 - \frac{1}{2}|u|^2\right).$$

The notation $f \propto g$ means that functions f and g coincide up to a constant, i.e., there is some $c > 0$ such that $f = cg$. The posterior is Gaussian and its mean and covariance, which can be found by completing the square, are given by

$$\theta_\delta = \frac{\langle A, m\rangle}{\delta^2 + \|A\|^2} \quad \text{and} \quad \sigma_\delta^2 = \frac{\delta^2}{\delta^2 + \|A\|^2}.$$

When the noise tends to zero we see that

$$\overline{\theta} = \lim_{\delta \to 0} \theta_\delta = \frac{\langle A, m_0\rangle}{\|A\|^2} \quad \text{and} \quad \overline{\sigma}^2 = \lim_{\delta^2 \to 0} \sigma_\delta^2 = 0.$$

The point $\overline{\theta}$ is the least-square solution for the linear equation $m = Au$. We see that the prior plays no role on the limit of zero observational noise.

Next we study the case $u \in \mathbb{R}^d$, $d \geq 2$, and $m \in \mathbb{R}$. The measurement is given by

$$m = \langle A, u\rangle + \eta,$$

with some $A \in \mathbb{R}^d \setminus \{0\}$. We assume that $\eta \sim \mathcal{N}(0, \delta^2)$ and $u \sim \mathcal{N}(0, \Sigma_0)$. Then

$$\pi^m(u) \propto \exp\left(- \frac{1}{2\delta^2}|m - \langle A, u\rangle|^2 - \frac{1}{2}\langle u, \Sigma_0^{-1}u\rangle\right).$$

We known that, as an exponential of a quadratic form, the posterior is a Gaussian measure with mean and covariance

$$\theta_\delta = \frac{m\Sigma_0 A}{\delta^2 + \langle A, \Sigma_0 A\rangle} \quad \text{and} \quad \Sigma_\delta = \Sigma_0 - \frac{(\Sigma_0 A)(\Sigma_0 A)^*}{\delta^2 + \langle A, \Sigma_0 A\rangle}.$$

When the noise tends to zero we get

$$\overline{\theta} = \lim_{\delta\to 0} \theta_\delta = \frac{m_0\Sigma_0 A}{\langle A, \Sigma_0 A\rangle} \quad \text{and} \quad \overline{\Sigma} = \lim_{\delta\to 0} \Sigma_\delta = \Sigma_0 - \frac{(\Sigma_0 A)(\Sigma_0 A)^*}{\langle A, \Sigma_0 A\rangle}.$$

We note that $\langle \overline{\theta}, A\rangle = m_0$ and $\overline{\Sigma}A = 0$. That is, when the observational noise decreases knowledge of u in the direction of A becomes certain. However, the uncertainty remains in directions not aligned with A. The magnitude of this uncertainty is determined by interaction between the properties of the prior and forward operator A. We see that in the underdetermined case the prior plays an important role even when the observational noise disappears.

Definition 1.12. *Let μ_n, $n \in \mathbb{N}$, and μ be two probability measures on $(X, \mathcal{B}(X))$. We say that μ_n converges weakly to μ if, for all bounded and continuous functions f, it holds that*

$$\lim_{n\to\infty} \int_X f(x) d\mu_n(x) = \int_X f(x) d\mu(x).$$

If this is the case, we write $\mu_n \rightharpoonup \mu$.

Lemma 1.13. *Let $\mu_n = \mathcal{N}(\theta_n, \Sigma_n)$ and $\mu = \mathcal{N}(\theta, \Sigma)$ on $\mathbb{R}^d$. If $\theta_n \to \theta$ and $\Sigma_n \to \Sigma$, as $n \to \infty$, then $\mu_n \rightharpoonup \mu$.*

Example 1.14. Let us return to the deblurring Example 1.1. In real life we only observe the signal m at finite number of observation points on a finite interval

$$m_i = m(t_i) = \int_0^1 a(t_i - s)u(s)ds + n(t_i), \quad 1 \le i \le k,$$

where we assume a to be of the form

$$a(t-s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(t-s)^2\Big).$$

We will also discretise the unknown u on the same mesh and approximate the integral as

$$\int_0^1 a(t_i - s)u(s)ds \approx \sum_{j=1}^k \frac{1}{k} a(t_i - s_j)u(s_j) = \sum_{j=1}^k a_{ij}u_j,$$

where we have denoted $s_j = \frac{j-1}{k-1}$, $u_j = u(s_j)$ and $a_{ij} = \frac{1}{k}a(t_i - s_j)$.

We have now discrete model $m = Au + n$, where $u, m, n \in \mathbb{R}^k$. To employ the Bayesian approach we will consider the stochastic model

$$m = Au + \eta$$

where m, η and u are treated as random variables. We assume that η is Gaussian noise with variance $\delta^2 I$,

$$\eta \sim \mathcal{N}(0, \delta^2 I) \quad \rho(\eta) \propto \exp\Big(-\frac{1}{2\delta^2} \|\eta\|^2 \Big).$$

Then the likelihood density is given as

$$\rho_u(m) = \rho(m - Au) \propto \exp\Big(-\frac{1}{2\delta^2} \|m - Au\|^2 \Big).$$

Next we have to choose a prior for the unknown. Assume that we know that $u(0) = u(1) = 0$ and u is quite smooth, that is, the value of $u(t)$ in a point is more or less the same as its neighbour. We will then model the unknown as

$$u_j = \frac{1}{2}(u_{j-1} + u_{j+1}) + W_j$$

where the innovative term W_j follows Gaussian distribution $\mathcal{N}(0, \gamma^2)$. The variance γ^2 determines how much the reconstructed function u departs from the smoothness model $u_j = \frac{1}{2}(u_{j-1} + u_{j+1})$. We can then write in matrix form

$$Lu = W, \quad \text{where} \quad L = \frac{1}{2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

Therefore the prior can be written as

$$\pi(u) \propto \exp\Big(-\frac{1}{2\gamma^2} \|Lu\|^2 \Big).$$

Using the Bayes' formula we get the posterior distribution

$$\pi^m(u) \propto \exp\Big(-\frac{1}{2\delta^2} \|m - Au\|^2 - \frac{1}{2\gamma^2} \|Lu\|^2 \Big).$$

In the next Section we will see how to extract useful information from the posterior distribution.

1.4 Estimators

The dimension of the inverse problem can be large and consequently the posterior distribution lives in a high dimensional space which makes its visualisation difficult. However, we can calculate different point estimators and spread or region estimators. The point estimators approximate the most probable value of the unknown given the data and the prior. The spread estimators give a region that contain the unknown with some high probability.

One of the most used statistical estimators is the maximum a posterior estimate (MAP), which is the mode of the posterior distribution. That is, given the posterior density $\pi^m(u)$ the MAP estimate u_{MAP} satisfies

$$u_{MAP} = \arg\max_{u \in \mathbb{R}^d} \pi^m(u),$$

if such maximiser exists. Note that MAP estimator may not be unique.

Another widely used point estimate is the conditional mean (CM) of the unknown u given the data m, which is defined by

$$u_{CM} = \mathbb{E}(u \,|\, m) = \int_{\mathbb{R}^d} u\pi(u \,|\, m)du,$$

provided that the integral converges. The main problem with CM estimation is that solving the integral in high-dimensional space is often very difficult.

As an example of spread estimate we can consider Bayesian credible sets. A level $1-\alpha$ credible set $\mathcal{C}_\alpha$, with some small $\alpha \in (0,1)$, is defined as

$$\Pi(\mathcal{C}_\alpha \,|\, m) = \int_{\mathcal{C}_\alpha} \pi(u \,|\, m)du = 1-\alpha.$$

Hence a credible set $\mathcal{C}_\alpha$ is a region that contains a large fraction of the posterior mass.

Another way of quantifying uncertainty is to consider problem $m^\dagger = Au^\dagger + \eta$, where $u^\dagger$ is though to be a deterministic 'true' unknown. We would then like to find random sets $\mathcal{C}_\alpha$ that frequently contain the 'true' unknown $u^\dagger$, that is, $\mathbb{P}(u^\dagger \in \mathcal{C}_\alpha) = 1-\alpha$. The set $\mathcal{C}_\alpha$ is called a frequentist confidence region of level $1-\alpha$.

Example 1.15. Let $u \in \mathbb{R}$ and assume that the posterior distribution is given by

$$\pi^m(u) = \frac{c}{\sigma_1}\phi\Big(\frac{u}{\sigma_1}\Big) + \frac{1-c}{\sigma_2}\phi\Big(\frac{u-1}{\sigma_2}\Big),$$

where $0 < c < 1$, $\sigma_1, \sigma_2 > 0$ and ϕ is density function of standard normal distribution $\phi(u) = (2\pi)^{-1/2}\exp(-u^2/2)$. In that case

$$u_{CM} = 1-c$$

and

$$u_{MAP} = \begin{cases} 0 & \text{if } c/\sigma_1 > (1-c)/\sigma_2, \\ 1 & \text{if } c/\sigma_1 < (1-c)/\sigma_2. \end{cases}$$

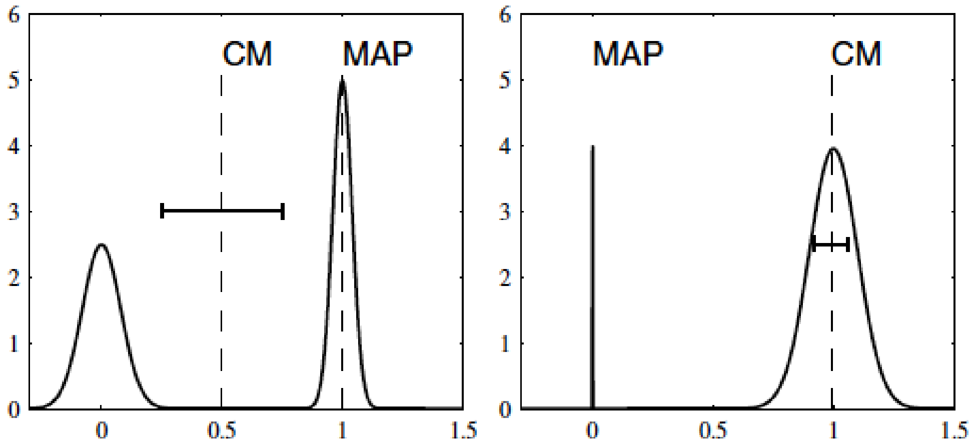

Figure 1: We can not say that one point estimator is better than the other in all applications. When CM gives a poor estimate the posterior has a larger variance.

If $c = 1/2$ and σ_1, σ_2 are small the probability that u takes values near u_{CM} is small. On the other hand if $\sigma_1 = c\sigma_2$ then $c/\sigma_1 > (1-c)\sigma_2$ and $u_{MAP} = 0$. But if c is small this is a bad estimate for u, since the likelihood for u to take value near 0 is less that c.

We can also calculate the posterior variance

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (u - u_{CM})^2 \pi^m(u) du \\ &= \int_{-\infty}^{\infty} u^2 \pi^m(u) du - u_{CM} \\ &= c\sigma_1^2 + (1-c)(\sigma_2^2 + 1) - (1-c)^2.\end{aligned}$$

We notice that when the conditional mean gives poor estimate the posterior variance is larger.

Example 1.16. Let us return to Example 1.14 where we got the posterior distribution

$$\pi^m(u) \propto \exp\Big(-\frac{1}{2\delta^2}\|m - Au\|^2 - \frac{1}{2\gamma^2}\|Lu\|^2\Big).$$

Since the posterior distribution is also Gaussian we know that the MAP and CM estimators coincides and we have an estimator

$$u^\delta_{MAP} = \arg\max_{u\in\mathbb{R}^k} \pi(u\,|\,m) = \arg\min_{u\in\mathbb{R}^k} \Big\{\frac{1}{2\delta^2}\|m - Au\|^2 + \frac{1}{2\gamma^2}\|Lu\|^2\Big\}.$$

Notice that u_{MAP} is of the same form as the Tikhonov estimator introduced in (1.4).

Completing the square we can write the posterior in form

$$\pi^m(u) \propto \exp\Big(-\frac{1}{2}\Big\|u - \frac{1}{\delta^2}\Gamma^{-1}A^\top m\Big\|^2_\Gamma\Big),$$

where we have used the weighted norm $\|\cdot\|_\Gamma = \|\Gamma^{\frac{1}{2}}\cdot\|$ with $\Gamma = \frac{1}{\delta^2}A^\top A + \frac{1}{\gamma^2}L^\top L$. Hence we see that the MAP estimator is given by

$$u_{MAP} = \frac{1}{\delta^2}\Gamma^{-1}A^\top m = \Big(A^\top A + \frac{\delta^2}{\gamma^2}L^\top L\Big)^{-1}A^\top m$$

and the posterior covariance is $\Sigma = \Gamma^{-1}$.

1.5 Prior models

Constructing a good prior density is one of the most challenging parts of solving a Bayesian inverse problem. The main problem is transforming our qualitative information into a quantitative form that can be coded as a prior density. The prior probability distribution should be concentrated on those values of u we expect to see and assigns a clearly higher probability to them than to the unexpected ones.

1.5.1 Gaussian prior

Gaussian probability densities are the most used priors in statistical inverse problems. They are easy to construct and form a versatile class of densities. They also often lead to explicit estimators. Due to the central limit theorem the Gaussian densities are often good approximation to inherently non-Gaussian distributions when the observation is based on a large number of mutually independent random events. This is also the reason why the noise is often assumed to be Gaussian.

Definition 1.17. *Let $\theta \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix. A Gaussian d-variate random variable u with mean θ and covariance Σ is a random variable with the probability density*

$$\pi(u) = \frac{1}{(2\pi|\Sigma|)^{d/2}}\exp\Big(-\frac{1}{2}(u-\theta)^\top\Sigma^{-1}(u-\theta)\Big),$$

where $|\Sigma| = \det(\Sigma)$. We then denote $u \sim \mathcal{N}(\theta,\Sigma)$.

The Gaussian distribution is completely characterised by its mean and covariance. Notice that the expression $(u-\theta)^\top\Sigma^{-1}(u-\theta)$ can also be written in form $\|\Sigma^{-1/2}u\|_2^2$, since due to our assumptions on Σ the inverse square root $\Sigma^{-1/2}$ is well-defined.

If we consider linear inverse problems and assume Gaussian prior and Gaussian noise model the posteriori distribution is of the form $c\cdot\exp(-G(u))$, where G can be rewritten as a sum of a quadratic form and constant term in order to show that the posterior is Gaussian. This method is called completing the square. In order to analyse the Gaussian posterior, we need some machinery from linear algebra.

Definition 1.18. *Let*

$$\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} \in \mathbb{R}^{d\times d}$$

be a positive definite symmetric matrix, where $\Sigma_{11} \in \mathbb{R}^{n\times n}$, $\Sigma_{22} \in \mathbb{R}^{(d-n)\times(d-n)}$, $n < d$ *and* $\Sigma_{21} = \Sigma_{12}^\top$. *We define the Schur complements* $\widetilde{\Sigma}_{jj}$ *of* Σ_{jj}, $j = 1, 2$, *as*

$$\widetilde{\Sigma}_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \quad \textit{and} \quad \widetilde{\Sigma}_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

The positive definiteness of Σ implies that Σ_{jj}, $j = 1, 2$, are also positive definite and hence the Schur complements are well defined. The following matrix inversion lemma is useful when calculating the conditional covariance.

Lemma 1.19. *Let* Σ *be a matrix satisfying the assumptions of Definition 1.18. Then the Schur complements* $\widetilde{\Sigma}_{jj}$ *are invertible matrices and*

$$\Sigma^{-1} = \begin{bmatrix} \widetilde{\Sigma}_{22}^{-1} & -\widetilde{\Sigma}_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\widetilde{\Sigma}_{11}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \widetilde{\Sigma}_{11}^{-1} \end{bmatrix}.$$

Let $u \sim \mathcal{N}(\theta_u, \Sigma_u)$ and $\eta \sim \mathcal{N}(0, \Sigma_\eta)$ with u and η independent. The distribution of $m = Au + \eta$ is Gaussian with $\theta_m = \mathbb{E}(m) = A\theta_u$ and

$$\Sigma_m = \mathbb{E}\big((m - \theta_m)(m - \theta_m)^\top\big) = A\Sigma_u A^\top + \Sigma_\eta.$$

We can also calculate

$$\mathbb{E}\big((u - \theta_u)(m - \theta_m)^\top\big) = \Sigma_u A^\top$$

The joint distribution of u and m then has a covariance

$$\operatorname{Cov}\begin{bmatrix} u \\ m \end{bmatrix} = \begin{bmatrix} \Sigma_u & \Sigma_u A^\top \\ A\Sigma_u & A\Sigma_u A^\top + \Sigma_\eta \end{bmatrix}$$

Hence the joint probability density of u and m is given by

$$\nu(u, m) \propto \exp\Big(-\frac{1}{2} \begin{bmatrix} u - \theta_u \\ m - \theta_m \end{bmatrix}^\top \begin{bmatrix} \Sigma_u & \Sigma_u A^\top \\ A\Sigma_u & A\Sigma_u A^\top + \Sigma_\eta \end{bmatrix}^{-1} \begin{bmatrix} u - \theta_u \\ m - \theta_m \end{bmatrix} \Big).$$

Theorem 1.20. *Assume that* $u : \Omega \to \mathbb{R}^d$ *and* $\eta : \Omega \to \mathbb{R}^k$ *are mutually independent Gaussian random variables*

$$u \sim \mathcal{N}(\theta_u, \Sigma_u), \quad \eta \sim \mathcal{N}(0, \Sigma_\eta),$$

where $\Sigma_u \in \mathbb{R}^{d\times d}$ *and* $\Sigma_\eta \in \mathbb{R}^{k\times k}$ *are positive definite. The noisy measurement* m *is given by* $m = Au + \eta$, *where* $A \in \mathbb{R}^{k\times d}$ *is a known matrix. Then the posterior probability density of* u *given the measurement* m *is*

$$\pi(u \,|\, m) \propto \exp\Big(-\frac{1}{2}(u - \overline{u})^\top \Sigma^{-1}(u - \overline{u})\Big),$$

where

$$\overline{u} = \theta_u + \Sigma_u A^\top (A\Sigma_u A^\top + \Sigma_\eta)^{-1}(m - A\theta_u)$$

and

$$\Sigma = \Sigma_u - \Sigma_u A^\top (A\Sigma_u A^\top + \Sigma_\eta)^{-1} A\Sigma_u.$$

Proof. By shifting the coordinate origin to $[\theta_u, \theta_m]$ we may assume that $\theta_u = \theta_m = 0$. By Bayes' formula we have $\pi(u \,|\, m) \propto \nu(u, m)$ and hence we will consider the joint density as a function of u. We denote the components of $\mathrm{Cov}([u\ m]^\top)$ by Σ_{ij}, $i, j = 1, 2$. Then, by Lemma 1.19 and the fact that $\Sigma_{22}^{-1}\Sigma_{21}\widetilde{\Sigma}_{22}^{-1} = \widetilde{\Sigma}_{11}^{-1}\Sigma_{21}\Sigma_{11}^{-1}$ (the covariance is symmetric), we have

$$\begin{aligned}\nu(u,m) &\propto \exp\Big(-\frac{1}{2}(u^\top\widetilde{\Sigma}_{22}^{-1}u - 2u^\top\widetilde{\Sigma}_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1}m + m^\top\widetilde{\Sigma}_{11}^{-1}m)\Big)\\ &= \exp\Big(-\frac{1}{2}(u - \Sigma_{12}\Sigma_{22}^{-1}m)^\top\widetilde{\Sigma}_{22}^{-1}(u - \Sigma_{12}\Sigma_{22}^{-1}m) + c\Big),\end{aligned}$$

where

$$c = m^\top(\widetilde{\Sigma}_{11}^{-1} - \Sigma_{22}^{-1}\Sigma_{21}\widetilde{\Sigma}_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1})m$$

is a constant independent of u and can hence be factored out of the density. □

Note that the posterior covariance is independent of the prior mean θ (and mean of the noise even if that would be non-zero). We have a more compact expression for the posterior mean and variance

Lemma 1.21. *Assume that u, η, m are as in Theorem 1.20. We then have*

$$\pi^m(u) \propto \exp\Big(-\frac{1}{2}(u - \overline{u})^\top\Sigma^{-1}(u - \overline{u})\Big),$$

where

$$\Sigma = (A^\top\Sigma_\eta^{-1}A + \Sigma_u^{-1})^{-1}$$

and

$$\overline{u} = \Sigma(A^\top\Sigma_\eta^{-1}m + \Sigma_u^{-1}\theta_u).$$

Proof left as an exercise.

Consider next a problem where the unknown is a two-dimensional pixel image. Let $u \in \mathbb{R}^d$ be the pixel image (which we have arranged as a vector), where a component u_j represents the intensity of the j^{th} pixel. Since we consider images it is natural to add a positivity constraint to our prior. Gaussian white noise density with positivity constraint is

$$\pi(u) \propto \mu_+(u)\exp\left(-\frac{1}{2\alpha^2}\|u\|_2^2\right)$$

where $\mu_+(u) = 1$ if $u_j > 0$ for all j, and $\mu_+(u) = 0$ otherwise. We assume that each component is independent of the others and hence the random draws can be performed componentwise. The one-dimensional distribution function can be defined as

$$\Phi(t) = \frac{1}{\alpha}\sqrt{\frac{2}{\pi}}\int_0^t \exp\left(-\frac{1}{2\alpha^2}s^2\right)ds = \mathrm{erf}\Big(\frac{t}{\alpha\sqrt{2}}\Big),$$

where erf stands for the error function

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t \exp(-s^2) ds.$$

The mutually independent components u_j are then drawn by

$$u_j = \Phi^{-1}(t_j) = \operatorname{erf}^{-1}\left(\frac{t_j}{\alpha\sqrt{2}} \right),$$

where t_js are drawn randomly from the uniform distribution $\mathcal{U}([0,1])$. The proof that this really produces draws from the prior is left as an exercise.

1.5.2 Impulse Prior

We assume again that the unknown is a two-dimensional pixel image. We have prior information is that the image contains small and well localised objects (for example a tumour in X-ray image). We can then use impulse prior model. These priors favour images with low average amplitude with few outliers. One example of such a prior is ℓ^1 prior. Let $u \in \mathbb{R}^d$ represent the pixel image, where a component u_j represents the intensity of the j^{th} pixel. The ℓ^1 prior is defined as

$$\pi(u) = \left(\frac{\alpha}{2}\right)^d \exp(-\alpha \|u\|_1),$$

where $\alpha > 0$ and $\|\cdot\|_1$ is the ℓ^1-norm. We can enhance the impulse noise effect by taking a smaller power of the components of u, that is, using $\sum |u_j|^p$, $p \in (0,1)$ instead of the ℓ^1-norm.

Another density that produces few distinctly different pixels with a low-amplitude background is the Cauchy density, which is defined as

$$\pi(u) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^{d} \frac{1}{1+\alpha^2 u_j^2}.$$

Let us take a closer look at the ℓ^1 prior and to what kind of draws it produces. Since we consider images we add a positivity constraint to our prior and write

$$\pi(u) = \alpha^d \mu_+(u) \exp(-\alpha \|u\|_1),$$

where $\mu_+(u) = 1$ if $u_j > 0$ for all j, and $\mu_+(u) = 0$ otherwise. The one-dimensional distribution function can be defined as

$$\Phi(t) = \alpha \int_0^t e^{-\alpha s} ds = 1 - e^{-\alpha t}.$$

The mutually independent components u_j are then drawn by

$$u_j = \Phi^{-1}(t_j) = -\frac{1}{\alpha} \log(1 - t_j),$$

where t_js are drawn randomly from the uniform distribution $\mathcal{U}([0,1])$. As mentioned before the proof that this really produces draws from the prior is left as an exercise.

Similarly when we draw independent components from the Cauchy distribution with positivity constraint we use the distribution function

$$\Phi(t) = \frac{2\alpha}{\pi}\int_0^t \frac{1}{1+\alpha^2 s^2}ds = \frac{2}{\pi}\arctan \alpha t$$

meaning that the inverse cumulative distribution is

$$\Phi^{-1}(t) = \frac{1}{\alpha}\tan\Big(\frac{\pi t}{2}\Big).$$

1.5.3 Discontinuities

Assume next that we want to estimate one-dimensional signal $f : [0,1] \to \mathbb{R}$, $f(0) = 0$, from indirect observations. Our prior knowledge is that the signal is usually relatively stable but can have large jumps every now and then. We may also have information on the size of the jumps or the rate of occurrence of the discontinuities. One possible prior is the finite difference approximation of the derivative of f with assumption that the derivative follows an impulse noise probability distribution. Let us discretise the interval $[0,1]$ by points $t_j = j/N$ and write $u_j = f(t_j)$. Consider the density

$$\pi(u) = \Big(\frac{\alpha}{\pi}\Big)^N \prod_{j=1}^N \frac{1}{1+\alpha^2(u_j - u_{j-1})^2}.$$

To draw from the above distribution let us define new random variables

$$x_j = u_j - u_{j-1}, \quad 1 \le j \le N.$$

The probability distribution of these variables is

$$\pi(x) = \Big(\frac{\alpha}{\pi}\Big)^N \prod_{j=1}^N \frac{1}{1+\alpha^2 x_j^2},$$

that is, they are independent of each other and can hence be drawn from the one-dimensional Cauchy density. Note that $u = [u_1, \cdots, u_N]^\top \in \mathbb{R}^N$ satisfies $u = Bx$, where $B \in \mathbb{R}^{N\times N}$ is a lower triangular matrix such that $B_{ij} = 1$ for $i \ge j$. The idea of the above prior can be generalised to higher dimensions which brings us to total variation prior.

We start by defining the concept of total variation for functions. Let $f : D \to \mathbb{R}$ be a function in $L^1(D)$, $D \subset \mathbb{R}^d$. We define the total variation of f, denoted by $\mathrm{TV}(f)$, as

$$\mathrm{TV}(f) = \sup_g \Big\{ \int_D f\nabla\cdot g\, dx \,|\, g = (g_1, \cdots, g_d) \in C_0^1(D, \mathbb{R}^d),\ \|g\|_{L^\infty} \le 1\Big\}.$$

The test function space $C_0^1(D, \mathbb{R}^d)$ consist of continuously differentiable vector-valued functions on D that vanish at the boundary. A function is said to have bounded variation if $\mathrm{TV}(f) < \infty$. To understand the definition let us consider the following simple example

Example 1.22. Let $D \subset \mathbb{R}^2$ be an open set and $B \subset D$ be a set bounded by a smooth curve $\partial B = S$, which does not intersect with the boundary of D. Let $f : D \to \mathbb{R}$ be the characteristic function of B. Let $g \in C_0^1(D, \mathbb{R}^2)$ be an arbitrary test function. By the divergence theorem we obtain

$$\int_D f \nabla \cdot g \, dx = \int_B \nabla \cdot g \, dx = \int_{\partial B} n \cdot g \, dS,$$

where n is the exterior unit normal vector of ∂B. This integral attains its maximum, under the constraint $\|g\|_{L^\infty} \leq 1$, if we set $n \cdot g = 1$ identically. Hence

$$\mathrm{TV}(f) = \mathrm{length}(\partial B).$$

Notice that the weak derivative of f is the Dirac delta of the boundary curve, which cannot be be presented by an integrable function. Therefore, the space of functions with bounded variation differs from the corresponding Sobolev space.

We will next consider two dimensional problem and define a discrete analogue for TV. Let $D \in \mathbb{R}^2$ be bounded and divided in d pixels. We define two pixels as neighbours if they share a common edge. The total variation of the discrete image $u = [u_1, \cdots, u_d]^\top$ is then defined

$$\mathrm{TV}(u) = \sum_{j=1}^{d} V_j(u), \quad V_j(u) = \frac{1}{2} \sum_{i \in \mathcal{N}_j} |u_i - u_j|,$$

where $\mathcal{N}_j$ is the neighbourhood of pixel u_j ($j \notin \mathcal{N}_j$). The discrete total variation density is then given by

$$\pi(u) \propto \exp(-\alpha \mathrm{TV}(u)).$$

The total variation density is concentrated on images that are 'blocky' consisting of blocks with short boundaries and small variation within each block.

The total variation prior is an example of a structural prior. Different structural priors, depending on different neighbourhood systems, can be derived from the theory of Markov random fields. For more prior choices and examples see e.g. [7, Section 3.3]

1.6 Sampling methods

An important part of Bayesian inversion techniques is to develop methods for exploring the posterior probability densities. We will next discus a random sampling methods known as the Markov chain Monte Carlo (MCMC) techniques. We saw previously in Section 1.4 that finding a MAP estimate leads to an optimisation problem, whereas the conditional mean requires integration over the space $\mathbb{R}^d$ where the posterior density is defined. Since the dimension of the problem can be large instead of calculating the full integral we want to sample from the posterior and then use these sample points to approximate the integral.

Let μ denote a probability measure on $\mathbb{R}^d$ and let f be a measurable function integrable over $\mathbb{R}^d$ with respect to μ, that is, $f \in L^1_\mu$. We want to estimate the integral of f with respect to the measure μ. In numerical quadrature methods one defines a set of support points $x_j \in \mathbb{R}^d$, $1 \leq j \leq N$ and the corresponding weights w_j to get an approximation

$$\int_{\mathbb{R}^d} f(x) d\mu(x) \approx \sum_{j=1}^{N} w_j f(x_j).$$

The above method is designed for computing one-dimensional integrals. To compute integrals in multiple dimensions, we could phrase the integral as repeated one-dimensional integrals by applying Fubini's theorem. However, this approach requires the function evaluations to grow exponentially as the number of dimensions increases which makes it infeasible in high dimensions.

In Monte Carlo integration the support points x_j are generated randomly by drawing from some probability density (ideally determined by μ) and the weights are then determined from the distribution μ. Assume that $x \sim \mu$. If we had a random generator such that repeated realisations of x could be produced we could generate a set of points distributed according to μ. We could then approximate the integral of f by the so called ergodic average,

$$\int_{\mathbb{R}^d} f(x) d\mu(x) = \mathbb{E}\big(f(x)\big) \approx \frac{1}{N} \sum_{j=1}^{N} f(x_j),$$

where $\{x_1, \cdots, x_N\} \subset \mathbb{R}^d$ is a representative collection of samples distributed according to μ.

The MCMC methods are systematic way of generating sample collection so that the above approximation holds. We start with some basic tools from probability theory

Definition 1.23. *A mapping $P : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ is called a probability transition kernel if*

1. *for each $B \in \mathcal{B}(\mathbb{R}^d)$ the mapping $\mathbb{R}^d \to [0, 1]$, $x \mapsto P(x, B)$ is a measurable function;*

2. *for each $x \in \mathbb{R}^d$ the mapping $\mathcal{B}(\mathbb{R}^d) \to [0, 1]$, $B \mapsto P(x, B)$ is a probability distribution.*

A discrete time stochastic process is an ordered set $\{x_j\}_{j=1}^{\infty}$ of random variables $x_j \in \mathbb{R}^d$. A time homogeneous Markov chain with the transition kernel P is a stochastic process $\{x_j\}_{j=1}^{\infty}$ with the properties

$$\mu_{x_{j+1}}(B_{j+1} \,|\, x_1, \cdots, x_j) = \mu_{x_{j+1}}(B_{j+1} \,|\, x_j) = P(x_j, B_{j+1}).$$

The first equality means that the probability $x_{j+1} \in B_{j+1}$ depends on the past only through the previous state x_j. The second equality states that time is homogeneous in the sense that the dependence of consecutive moments does not vary in time since the kernel P does not depend on time j.

We define the transition kernel that propagates k steps forward as

$$P^k(x_j, B_{j+k}) = \mu_{x_{j+k}}(B_{j+k} \,|\, x_j) = \int_{\mathbb{R}^d} P(x_{j+k-1}, B_{j+k}) P^{k-1}(x_j, dx_{j+k-1}), \quad k \geq 2,$$

where $P^1(x_j, B_{j+1}) = P(x_j, B_{j+1})$. if μ_{x_j} denotes the probability distribution of x_j the distribution of x_{j+1} is given by

$$\mu_{x_{j+1}}(B_{j+1}) = \mu_{x_j} P(B_{j+1}) = \int_{\mathbb{R}^d} P(x_j, B_{j+1}) d\mu_{x_j}(x_j).$$

We will next introduce few concepts concerning the transition kernels

1. The measure μ is an *invariant measure* of $P(x_j, B_{j+1})$ if

   $$\mu P = \mu.$$

   This means that the distribution of the random variable x_j before the time step $j \to j+1$ is the same as the variable x_{j+1} after the step.

2. The transition kernel P is *irreducible* (with respect to a given measure μ) if for each $x \in \mathbb{R}^d$ and $B \in \mathcal{B}(\mathbb{R}^d)$, with $\mu(B) > 0$, there exists an integer k such that $P^k(x, B) > 0$. This means that regardless of the starting point the Markov chain generated by P visits any set of positive measure with positive probability.

3. Let P be irreducible kernel. We say that P is *periodic* if, for some integer $m \geq 2$, there is a set of disjoint non-empty sets $\{E_1, \cdots, E_m\} \subset \mathbb{R}^d$ such that $P(x, E_{j+1(\bmod m)}) = 1$ for all $j = 1, \cdots, m$ and all $x \in E_j$. This means that a periodic P generates a Markov chain that remains in a periodic loop for ever. We say that P is an *aperiodic* kernel if it is not periodic.

The following theorem is of crucial importance for MCMC methods.

Theorem 1.24. *Let μ be a probability measure on $\mathbb{R}^d$ and $\{x_j\}$ a time homogeneous Markov chain with transition kernel P. Assume further that μ is an invariant measure of the transition kernel, and that P is irreducible and aperiodic. Then for all $x \in \mathbb{R}^d$,*

$$\lim_{N \to \infty} P^N(x, B) = \mu(B), \quad \textit{for all } B \in \mathcal{B}(\mathbb{R}^d),$$

and for $f \in L^1_\mu(\mathbb{R}^d)$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} f(x_j) = \int_{\mathbb{R}^d} f(x) d\mu(x)$$

almost certainly.

1.6.1 Metropolis-Hastings

Let Π denote the target probability distribution in $\mathbb{R}^d$. We assume that Π is absolutely continuous with respect to Lebesgue measure and has density $\pi(x)$. We want to determine a transition kernel $P(x, B)$ so that Π is its invariant measure.

Let P denote any transition kernel. If we start from a point $x \in \mathbb{R}^d$ the kernel either proposes to move to another point $y \in \mathbb{R}^d$ or to stay in x. Hence we can split the kernel in two parts,

$$P(x, B) = \int_B K(x, y)dy + r(x)\chi_B(x),$$

where χ_B is the characteristic function of $B \in \mathcal{B}(\mathbb{R}^d)$. Loosely speaking $K(x, y) \geq 0$ describes the probability for moving and $r(x) \geq 0$ the probability for staying put.

The condition $P(x, \mathbb{R}^d) = 1$ implies that

$$r(x) = 1 - \int_{\mathbb{R}^d} K(x, y)dy. \tag{1.8}$$

We assume that the K satisfies the *detailed balance condition*

$$\pi(y)K(y, x) = \pi(x)K(x, y), \tag{1.9}$$

for all $x, y \in \mathbb{R}^d$. This guarantees that Π is an invariant measure of P since using (1.8) we can then write

$$\begin{aligned} \Pi P(B) &= \int_{\mathbb{R}^d} \Big(\int_B K(x, y)dy + r(x)\chi_B(x) \Big) \pi(x)dx \\ &= \int_B \Big(\int_{\mathbb{R}^d} \pi(x)K(x, y)dx + r(y)\pi(y) \Big) dy \\ &= \int_B \pi(y)dy \end{aligned}$$

Our goal now is to construct a transition kernel that K that satisfies the detailed balance equation 1.9. Let $q : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a given functional with property $\int q(x, y)dy = 1$. The function q is called the *proposal distribution* and it defines a transition kernel

$$Q(x, A) = \int_A q(x, y)dy.$$

If q satisfies the detailed balance condition we can simply choose $K(x, y) = q(x, y)$ and $r(x) = 0$. Otherwise we have to correct the kernel and define

$$K(x, y) = \alpha(x, y)q(x, y), \tag{1.10}$$

where α is a correction term.

Assume that instead of the detailed balance condition we have

$$\pi(y)q(y, x) < \pi(x)q(x, y),$$

for some $x, y \in \mathbb{R}^d$. Our aim is to choose α so that

$$\pi(y)\alpha(y,x)q(y,x) = \pi(x)\alpha(x,y)q(x,y).$$

We can achieve this by setting

$$\alpha(y,x) = 1 \quad \text{and} \quad \alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} < 1.$$

We then see that the kernel K defined in (1.10) satisfies the detailed balance condition (1.9) if we define

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right).$$

This transition kernel is called the Metropolis-Hastings kernel.

We can implement the method derived above using an algorithm that is carried out through the following steps;

1. Pick initial value $x_1 \in \mathbb{R}^d$ and set $j = 1$.
2. Draw $y \in \mathbb{R}^d$ from the proposal kernel $q(x_j, y)$ and calculate the acceptance ratio
$$\alpha(x_j, y) = \min\left(1, \frac{\pi(y)q(y,x_j)}{\pi(x_j)q(x_j,y)}\right).$$
3. Draw $t \in [0,1]$ from uniform probability density.
4. If $t \leq \alpha(x_j, y)$, set $x_{j+1} = y$, otherwise $x_{j+1} = x_j$. Increase $j \to j+1$ and go to step 2. until $j = J$, the desired sample size.

Note that if the candidate generating the kernel is symmetric $q(x,y) = q(y,x)$ for all $x, y \in \mathbb{R}^d$ then the acceptance ration simplifies to

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right).$$

This means that we accept immediately moves towards higher probability and sometimes also moves that take us to lower probability.

Example 1.25. Consider a two-dimensional density

$$\pi(x) \propto \exp\left(- 10(x_1^2 - x_2)^2 - \left(x_2 - \frac{1}{4}\right)^4\right).$$

In what follows, we assume to have random number generators for $W \sim \mathcal{N}(0,1)$ and $t \sim \mathcal{U}([0,1])$ at our disposal (in Matlab the command randn and rand respectively).

We construct Metropolis–Hastings sequence using the random walk proposal distribution. We define

$$q(x,y) = \exp\left(- \frac{1}{2\gamma^2}\|x-y\|^2\right).$$

This means that we assume that the scaled random step from x to y is distributed as white noise $W = (y-x)/\gamma \sim \mathcal{N}(0,I)$. Using the above proposal distribution we get the following updating algorithm;

Algorithm 1: Simple Metropolis–Hastings update scheme

```
Pick initial value $x_1$. Set $x = x_1$;
for $j = 2 : J$ do
    Calculate $\pi(x)$;
    Draw $W \sim \mathcal{N}(0, I)$, set $y = x + \gamma W$;
    Calculate $\pi(y)$;
    Calculate $\alpha(x, y) = \min(1, \pi(y)/\pi(x))$;
    Draw $t \sim \mathcal{U}([0, 1])$;
    if $t < \alpha(x, y)$ then
        Accept: Set $x = y$, $x_j = x$;
        else
            Reject: Set $x_j = x$
        end if
    end if
end for
```

1.6.2 Single component Gibbs sampler

Gibbs sampling is used to sample multivariate distributions. The proposal kernel is defined using the density π to sample each component x_i of the vector $x = (x_1, \cdots, x_d)$ from the distribution of that component conditioned on all other components sampled so far.

If x is a d-variate random variable with the probability density π the probability density of the i^{th} component x_i conditioned on all x_j, for which $i \neq j$, is given by

$$\pi(x_i \,|\, x_{-i}) = C_i \pi(x)$$

where $x_{-i} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d)$ and C_i is a normalisation constant. We can then define a transition kernel K as

$$K(x, y) = \prod_{i=1}^{d} \pi(y_i \,|\, y_1, \cdots, y_{i-1}, x_{i+1}, \cdots, x_d)$$

and set $r(x) = 0$. This kernel does not usually satisfy the detailed balance condition but it satisfies a weaker but sufficient balance condition $\int_{\mathbb{R}^d} \pi(y) K(y, x) dx = \int_{\mathbb{R}^d} \pi(x) K(x, y) dx$.

The steps needed for implementing the algorithm can be summarised as follows;

1. Pick an initial value $x^1 \in \mathbb{R}^d$ and set $j = 1$.
2. Set $x = x^j$. For $1 \leq i \leq d$, draw $y_i \in \mathbb{R}$ from the one-dimensional distribution $\pi(y_i \,|\, y_1, \cdots, y_{i-1}, x_{i+1}, \cdots, x_d)$.
3. Set $x^{j+1} = y$. Increase $j \to j + 1$ and repeat from step 2. until j reaches the desired sample size J.

Example 1.26. We want to sample from the two-dimensional distribution

$$\pi(x) = \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \ \rho > 0.$$

In order to sample from this distribution using Gibbs sampler, we need to calculate the conditional distributions for directions x_1, x_2. We see that

$$\pi(x_1^j \mid x_2^{(j-1)}) = \mathcal{N}(\rho x_2^{(j-1)}, \sqrt{1-\rho^2}) \quad \text{and} \quad \pi(x_2^j \mid x_1^j) = \mathcal{N}(\rho x_1^j, \sqrt{1-\rho^2}).$$

We can write the algorithm as follows;

Algorithm 2: Simple Gibbs sampler update scheme

```
Pick initial value $x^1$. Set $x = x^1$;
for $j = 2 : J$ do
    Draw $x_1 \sim \mathcal{N}(\rho x_2 \sqrt{1-\rho^2})$;
    Draw $x_2 \sim \mathcal{N}(\rho x_1, \sqrt{1-\rho^2})$;
    Set $x^j = x$
end for
```

1.7 Hierarchical models

The prior densities we use depend on some parameters, such as variance or mean. So far we have assumed that these parameters are known. However, we often do not know how to choose them. If a parameter is not know, it can be estimated as a part of the statistical inference problem based on the data. This leads to hierarchical models that include hypermodels for the parameters defining the prior density.

Assume that the prior distribution depends on a parameter α, which is assumed to be unknown. We then write the prior as a conditional density

$$\pi(u \mid \alpha).$$

We model the unknown α with a hyperprior $\pi_h(\alpha)$ and write the joint distribution of u and α as

$$\pi(u, \alpha) = \pi(u \mid \alpha)\pi_h(\alpha).$$

Assuming we have a likelihood model $\pi(m \mid u)$ for the measurement m, we get the posterior density for u and α given m using the Bayes' formula

$$\pi(u, \alpha \mid m) \propto \pi(m \mid u, \alpha)\pi(u, \alpha) = \pi(m \mid u, \alpha)\pi(u \mid \alpha)\pi_h(\alpha).$$

The hyperprior density π_h may depend on some hyperparameter α_0. The main reason for the use of a hyperprior model is that the construction of the posterior is assumed to be more robust with respect to fixing a value for the hyperparameter α_0 than fixing a value for α. Sometimes we might want to treat also α_0 as a random variable with a respective probability density. We can then write $\pi(\alpha \mid \alpha_0)$ which leads to nested hypermodels.

Example 1.27. We return to the deblurring example 1.14, where we assumed prior $\pi(u) \propto \exp\big(-\|Lu\|^2/2\theta\big)$, with L being the second order finite difference matrix and $\theta = \gamma^2$ was assumed to be known.

We will next assume that we do not know the value of θ. The prior for u given θ is

$$\pi(u\,|\,\theta) = C_\theta \exp\Big(-\frac{1}{2\theta}\|Lu\|^2\Big).$$

The integral of a density is 1 and hence, with the change of variables $u = \sqrt{\theta}z$, $du = \theta^{d/2}dz$, we see that

$$1 = C_\theta \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2\theta}\|Lu\|^2\Big)du = \theta^{d/2}C_\theta \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\|Lz\|^2\Big)dz.$$

The last integral does not depend on θ so we deduce that $C_\theta \propto \theta^{-d/2}$ and write

$$\pi(u\,|\,\theta) \propto \exp\Big(-\frac{1}{2\theta}\|Lu\|^2 - \frac{d}{2}\log(\theta)\Big).$$

Since θ is not known we will treat it as a random variable. Any information concerning θ is then coded in the prior probability density π_h. The inverse problem is to approximate pair of unknowns (u, θ). If we only know that $\theta > 0$ we can use the improper prior density

$$\pi_h(\theta) = \pi_+(\theta) = \begin{cases} 0 & \text{if } \theta < 0, \\ 1 & \text{if } \theta \geq 0. \end{cases}$$

Note that π_+ is an improper density, since it is not integrable. In practice, we can assume an upper bound that we hope will never play a role. The posterior density is then given as

$$\pi^m(u,\theta) = \pi_+(\theta)\exp\Bigg(-\frac{1}{2\delta^2}\|m - Au\|^2 - \frac{1}{2\theta}\|Lu\|^2 - \frac{d}{2}\log(\theta)\Bigg).$$

To find the MAP estimator we can use sequential optimisation, where we update the value for u using the value for θ from previous step and then use this value of u to update θ;

1. Initialise $\theta = \theta_0$, set $k = 1$.

2. Update u,

$$u^k = \arg\max_{u\in\mathbb{R}^d}\{\pi^m(u\,|\,\theta^{k-1})\} = \arg\min_{u\in\mathbb{R}^d}\left\{\frac{1}{2\delta^2}\|m - Au\|^2 + \frac{1}{2\theta}\|Lu\|^2\right\}.$$

3. Update θ,

$$\theta^k = \arg\max_{\theta\geq 0}\{\pi^m(\theta\,|\,u^k)\} = \arg\min_{\theta\geq 0}\left\{\frac{1}{2\theta}\|Lu\|^2 + \frac{d}{2}\log(\theta)\right\}.$$

4. Increase k by one and repeat from 2. until convergence.

We calculated the update for u in Example 1.16. For θ we notice that the derivative is zero in the minimum of the function, that is,

$$-\frac{1}{\theta^2}\|Lu\|^2 + \frac{d}{\theta} = 0 \quad \Rightarrow \quad \theta = \frac{\|Lu\|^2}{d}.$$

Assume next that we know that the signal varies slowly except for unknown number of jumps of unknown size and location. The jumps should be sudden, suggesting that the variances should be mutually independent. This means that instead of assuming $W_j \sim \mathcal{N}(0,\theta)$ we should assume $W_j \sim \mathcal{N}(0,\theta_j)$. Then

$$\pi(u\,|\,\theta) \propto \exp\bigg(-\frac{1}{2}\|D_\theta^{-1/2}Lu\|^2 - \frac{1}{2}\sum_{j=1}^{d}\log(\theta_j)\bigg),$$

where $D_\theta = \mathrm{diag}(\theta)$. Only a few variances can be significantly large, while most of them should be small, suggesting a hyperprior that allows rare outliers.

One option is to use Gamma distribution as a prior for θ_j

$$\theta_j \sim \mathrm{Gamma}(\alpha,\theta_0), \quad \pi_h(\theta_j) \propto \prod_{j=1}^{d}\theta_j^{\alpha-1}\exp\bigg(-\frac{\theta_j}{\theta_0}\bigg).$$

Then, if $F(u,\theta\,|\,m) = -\log(\pi(u,\theta\,|\,m))$, we see that

$$F(u,\theta\,|\,m) \propto \frac{1}{2\delta^2}\|Au-m\|^2 + \frac{1}{2}\|D_\theta^{-1/2}Lu\|^2 + \frac{1}{\theta_0}\sum_{j=1}^{d}\theta_j - \Big(\alpha - \frac{3}{2}\Big)\sum_{j=1}^{d}\log(\theta_j).$$

We then get the following update step for u

$$u^k = \arg\min_{u\in\mathbb{R}^d}\bigg(\frac{1}{2\delta^2}\|Au-m\|^2 + \frac{1}{2}\|D_{\theta^{k-1}}^{-1/2}Lu\|^2\bigg).$$

To update θ we notice that θ_j^k satisfies

$$\frac{\partial}{\partial\theta_j}F(u^k,\theta) = -\frac{1}{2}\bigg(\frac{(Lu^k)_j}{\theta_j}\bigg)^2 + \frac{1}{\theta_0} - \Big(\alpha - \frac{3}{2}\Big)\frac{1}{\theta_j} = 0,$$

which has explicit solution

$$\theta_j^k = \theta_0\bigg(a + \sqrt{\frac{(Lu^k)_j^2}{2\theta_0} + a^2}\bigg), \quad a = \frac{1}{2}\Big(\alpha - \frac{3}{2}\Big).$$

2 Bayes' theorem on separable Banach spaces

In this section we prove a version of Bayes' theorem that can be used when the likelihood and prior are measures on separable Banach spaces. Note that there is no equivalent to Lebesgue measure in infinite dimensions (as it could not be σ-additive), and so we cannot define a measure by prescribing the form of its density. In our setting the posterior will always be absolutely continuous with respect to the prior. It is possible to construct examples even in purely Gaussian setting where this is not true. Hence working under this assumption is not strictly necessary but it is quite natural. Absolute continuity ensures that almost sure properties of the prior will be inherited by the posterior. To change such properties by data the data would have to contain infinite amount of information, which is unnatural in most applications. We follow the program laid out in [4] and [9].

Let X and Y denote measurable spaces and let ν and μ be probability measures on $X \times Y$. We assume that $\nu \ll \mu$. Using Theorem 1.3 we know that there exist a μ-measurable function $\phi : X \times Y \to \mathbb{R}$ with $\phi \in L^1_\mu$ such that

$$\frac{d\nu}{d\mu}(x,y) = \phi(x,y).$$

Theorem 2.1. *Assume that the conditional random variable $x \,|\, y$ exists under μ with probability distribution denoted by $\mu^y(dx)$. Then the conditional random variable $x \,|\, y$ under ν exists with probability distribution denoted by $\nu^y(dx)$. Furthermore $\nu^y \ll \mu^y$ and if $z(y) = \int_X \phi(x,y) d\mu^y(x) > 0$ we can write*

$$\frac{d\nu^y}{d\mu^y}(x) = \frac{1}{z(y)}\phi(x,y).$$

We will proceed to use the above theorem to construct the conditional distribution of the unknown u given data m from their joint probability distribution. We will need the following lemma to establish the measurability of the likelihood.

Lemma 2.2. *Let X be a Borel measurable topological space and assume that $g \in C(X;\mathbb{R})$, and that $\mu(X) = 1$ for some probability measure μ on X. Then g is a μ measurable function.*

2.1 Bayes' theorem for inverse problems

Let X, $\widetilde{Y}$ and Y be separable Banach spaces, equipped with the Borel σ-algebra, and let $A : X \to \widetilde{Y} \subset Y$ be a measurable linear mapping. We are interested in approximating u from a measurement

$$m = Au + \eta,$$

where $\eta \in Y$ denotes noise. We assume $(u, m) \in X \times Y$ to be a random variable and want to compute $u \,|\, m$. The random variable (u, m) is specified as follows:

- **Prior:** $u \sim \Pi$ measure on X.

- **Noise:** $\eta \sim P_0$ measure on Y and $\eta \perp u$.

The random variable $m \,|\, u$ is then distributed according to measure P_u, the translation of P_0 by Au. In the following we assume that $P_u \ll P_0$ for Π-a.s. Thus there exists a potential $\Phi : X \times Y \to \mathbb{R}$

$$\frac{dP_u}{dP_0}(m) = \exp(-\Phi(u; m)).$$

The mapping $\Phi(u; \cdot) : Y \to \mathbb{R}$ is measurable for a fixed u and $\mathbb{E}^{P_0} \exp(-\Phi(u; m)) = 1$. For a given realisation m of the data the function $-\Phi(\cdot; m)$ is called the *log likelihood.*

We define Q_0 to be the product measure

$$Q_0(du, dm) = \Pi(du)P_0(dm).$$

We assume that $\Phi(\cdot, \cdot)$ is Q_0 measurable. Then the random variable $(u, m) \in X \times Y$ is distributed according to measure $Q(du, dm) = \Pi(du)P_u(dm)$ and $Q \ll Q_0$ with

$$\frac{dQ}{dQ_0}(u, m) = \exp(-\Phi(u; m)).$$

We have the following infinite dimensional version of Theorem 1.9.

Theorem 2.3 (Bayes' Theorem)**.** *Assume that $\Phi : X \times Y \to \mathbb{R}$ is Q_0 measurable and that*

$$Z(m) = \int_X \exp(-\Phi(u; m))d\Pi(u) > 0$$

for P_0-a.s. Then the conditional distribution of $u \,|\, m$ exists under Q and is denoted by Π^m. Furthermore $\Pi^m \ll \Pi$ and

$$\frac{d\Pi^m}{d\Pi}(u) = \frac{1}{Z(m)} \exp(-\Phi(u; m)),$$

for m Q-a.s.

Proof. The positivity of $Z(m)$ holds Q_0 almost surely, and hence by the absolutely continuity of Q with respect to Q_0, it also holds Q almost surely. We can then use Theorem 2.1. Note that since $\mu = Q_0(du, dm)$ has a product form the conditional distribution of $u \,|\, m$ under Q_0 is simply Π. □

2.2 Well-posedness

In inverse problems small changes in data can cause large changes in the solution and hence some form of regularisation is needed to stabilise the problems. We will next show that Bayesian approach can be used to combat the ill-posedness of inverse problems so that small changes in the data will lead to small changes in the posterior measure.

In order to measure the changes in the posterior measure Π^m caused by the changes in the data we need a metric in measures. Let μ and μ' be probability measures on separable Banach space X and assume that they are both absolutely continuous with respect some reference measure ν defined in the same measure space (we can take for example $\nu = 1/2(\mu + \mu')$).

Definition 2.4. *We define the total variation distance between μ and μ' as*

$$d_{TV}(\mu, \mu') = \frac{1}{2} \int \left| \frac{d\mu}{d\nu} - \frac{d\mu'}{d\nu} \right| d\nu$$

If $\mu' \ll \mu$ we can simplify the above and write

$$d_{TV}(\mu, \mu') = \frac{1}{2} \int \left| 1 - \frac{d\mu'}{d\mu} \right| d\mu.$$

Definition 2.5. *We define the Hellinger distance between μ and μ' as*

$$d_{Hell}(\mu, \mu') = \sqrt{\frac{1}{2} \int \left(\sqrt{\frac{d\mu}{d\nu}} - \sqrt{\frac{d\mu'}{d\nu}} \right)^2 d\nu}$$

If $\mu' \ll \mu$ we can simplify the above and write

$$d_{Hell}(\mu, \mu') = \sqrt{\frac{1}{2} \left(1 - \sqrt{\frac{d\mu'}{d\mu}} \right)^2 d\mu}.$$

Note that we have

$$0 \leq d_{TV}(\mu, \mu') \leq 1 \quad \text{and} \quad 0 \leq d_{Hell}(\mu, \mu') \leq 1$$

Hellinger and total variation distances generate the same topology and we have the following inequalities.

Lemma 2.6. *The total variation and Hellinger metrics are related by the inequalities*

$$\frac{1}{\sqrt{2}} d_{TV}(\mu, \mu') \leq d_{Hell}(\mu, \mu') \leq \sqrt{d_{TV}(\mu, \mu')}.$$

Let X and Y be separable Banach spaces, equipped with the Borel σ-algebra, and let Π be a measure on X. We want to study the posterior distribution defined in the previous section. To make sense of it we need the following assumption.

Assumption 2.7. Let $X' \subset X$ and assume that $\Phi \in C(X' \times Y; \mathbb{R})$. Assume further that there are functions $M_i : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$, $i = 1, 2$, which are monotonic non-decreasing separately in each argument, and with M_2 strictly positive, such that for all $u \in X'$ and $m, m' \in B_Y(0, r)$,

$$\begin{aligned} -\Phi(u; m) &\leq M_1(r, \|u\|_X), \\ |\Phi(u; m) - \Phi(u; m')| &\leq M_2(r, \|u\|_X) \|m - m'\|. \end{aligned}$$

Theorem 2.8. *Let Assumption 2.7 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(X' \cap A) > 0$ *for some bounded set* $A \subset X$*. We also assume that*

$$\exp(M_1(r, \|u\|_X)) \in L^1_\Pi(X; \mathbb{R}), \tag{2.1}$$

for every fixed $r > 0$*. Then* $Z(m) = \int_X \exp(-\Phi(u; m))d\Pi(u)$ *is positive and finite for every* $m \in Y$ *and the posterior probability measure* Π^m *given by Theorem 2.3 is well defined.*

Proof. The boundedness of $Z(m)$ follows directly from the lower bound on Φ in Assumption 2.7 together with the integrability condition assumed in the theorem.

If $u \sim \Pi$ then $u \in X'$ a.s. and we can write

$$Z(m) = \int_{X'} \exp(-\Phi(u; m))d\Pi(u).$$

We also note that, since $A' = A \cap X'$ is bounded by assumption, $\sup_{u \in A'} \|u\|_X = R_1 < \infty$. Since $\Phi : X' \times Y \to \mathbb{R}$ is continuous it is finite at every point in $A' \times \{m\}$. Thus we see that

$$\sup_{(u,m) \in A' \times B_Y(0,r)} \Phi(u; m) = R_2 < \infty.$$

Hence

$$Z(m) \geq \int_{A'} \exp(-R_2)d\Pi(u) = \exp(-R_2)\Pi(A') > 0.$$

□

The above theorem shows that the measure Π^m is well-defined and normalisable. We did not need to check normalisability in Theorem 2.3 because Π^m was defined as a regular conditional probability via Theorem 2.1 which makes it automatically normalisable.

Theorem 2.9. *Let Assumption 2.7 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(A \cap X') > 0$ *for some bounded set* A *in* X*. We assume also that*

$$\exp(M_1(r, \|u\|_X))\Big(1 + M_2(r, \|u\|_X)^2\Big) \in L^1_\Pi(X; \mathbb{R}),$$

for every fixed $r > 0$*. Then there exists* $c = c(r) > 0$ *such that*

$$d_{Hell}(\Pi^m, \Pi^{m'}) \leq c\|m - m'\|_Y,$$

for all $m, m' \in B_Y(0, r)$*.*

Proof. Let $Z(m)$ and $Z(m')$ denote the normalisation constants for Π^m and $\Pi^{m'}$ so that

$$Z(m) = \int_{X'} \exp(-\Phi(u; m))d\Pi(u) > 0 \quad \text{and}$$
$$Z(m') = \int_{X'} \exp(-\Phi(u; m'))d\Pi(u) > 0$$

by Theorem 2.8. Using the local Lipschitz property of the exponential and the assumed Lipschitz continuity of $\Phi(u;\cdot)$ together with fact that $M_2(r,\|u\|_X) \leq 1 + M_2(r,\|u\|_X)^2$ we get

$$\begin{aligned}|Z(m)-Z(m')| &\leq \int_{X'} |\exp(-\Phi(u;m)) - \exp(-\Phi(u;m'))| d\Pi(u)\\ &\leq \int_{X'} \exp(M_1(r,\|u\|_X))|\Phi(u;m)-\Phi(u;m')|d\Pi(u)\\ &\leq \Big(\int_{X'} \exp(M_1(r,\|u\|_X))M_2(r,\|u\|_X)d\Pi(u)\Big)\|m-m'\|_Y\\ &\leq \Big(\int_{X'} \exp(M_1(r,\|u\|_X))\Big(1+M_2(r,\|u\|_X)\Big)d\Pi(u)\Big)\|m-m'\|_Y\\ &\leq c\|m-m'\|_Y.\end{aligned}$$

We use $c=c(r)$ to denote a constant independent of u and the value may change from occurrence to occurrence.

Using the definition of Hellinger distance and the fact that $(ab-cd)^2 \leq 2a^2(b-d)^2 + 2(a-c)^2d^2$ we get

$$\begin{aligned}\Big(d_{Hell}(\Pi^m, \Pi m')\Big)^2 &= \int_X \Big(Z(m)^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\Phi(u;m)\Big) - Z(m')^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\Phi(u;m')\Big)\Big)^2 d\Pi(u)\\ &\leq I_1 + I_2,\end{aligned}$$

where

$$\begin{aligned}I_1 &= \frac{1}{Z(m)}\int_{X'}\Big(\exp\Big(-\frac{1}{2}\Phi(u;m)\Big) - \exp\Big(-\frac{1}{2}\Phi(u;m')\Big)\Big)^2 d\Pi(u) \quad \text{and}\\ I_2 &= (Z(m)^{-\frac{1}{2}} - Z(m')^{-\frac{1}{2}})^2 \int_{X'} \exp\big(-\Phi(u;m')\big)d\Pi(u).\end{aligned}$$

Using the Assumption 2.7 and the fact that $Z(m) > 0$ we can use similar Lipschitz calculation as before and write

$$\begin{aligned}I_1 &\leq \frac{1}{4Z(m)}\int_{X'} \exp\big(M_1(r,\|u\|_X)\big)|\Phi(u;m)-\Phi(u;m')|^2 d\Pi(u)\\ &\leq \frac{\|m-m'\|_Y^2}{4Z(m)}\int_{X'} \exp\big(M_1(r,\|u\|_X)\big)M_2(r,\|u\|_X)^2 d\Pi(u)\\ &\leq c\|m-m'\|_Y^2.\end{aligned}$$

We note that Assumption 2.7 with (2.1) implies

$$\int_{X'} \exp\big(-\Phi(u;m')\big)d\Pi(u) \leq \int_{X'} \exp\big(M_1(r,\|u\|_X)\big)d\Pi(u) < \infty.$$

Hence

$$I_2 \leq \frac{c\big(Z(m)-Z(m')\big)^2}{\min\big(Z(m)^3, Z(m')^3\big)} \leq c\|m-m'\|_Y^2,$$

which completes the proof. □

Hellinger distance has the desirable property of giving bounds for expectations.

Lemma 2.10. *Let μ and μ' be two probability measures on a separable Banach space X. Assume that $f : X \to E$, where $(E, \|\cdot\|)$ is a separable Banach space, is measurable and has second moments with respect to both μ and μ'. Then*

$$\|\mathbb{E}^{\mu} f - \mathbb{E}^{\mu'} f\| \leq 2\sqrt{\mathbb{E}^{\mu}\|f\|^2 + \mathbb{E}^{\mu'}\|f\|^2}\; d_{Hell}(\mu, \mu')$$

The proof of the above lemma is left as an exercise.

Using Lemma 2.10 we see that, for $m, m' \in B_Y(0, r)$,

$$|\mathbb{E}^{\Pi^m} f(u) - \mathbb{E}^{\Pi^{m'}} f(u)| \leq c_{f,r}\|m - m'\|_Y$$

If Π is Gaussian we can use the following Fernique theorem to establish the integrability conditions in the above theorems.

Theorem 2.11 (Fernique)**.** *Let Π be a Gaussian probability measure on a separable Banach space X. Then there exists $\alpha > 0$ such that*

$$\int_X \exp(\alpha\|u\|_X^2) d\Pi(u) < \infty.$$

2.3 Approximation of the potential

In this section we will examine the continuity properties of the posterior measure with respect to approximation of the potential Φ. The data m is assumed to be fixed in this section so we will write $Z(m) = Z$ and $\Phi(u; m) = \Phi(u)$. Let X be a separable Banach space and Π a measure on X. Assume that Π^m and Π^m_N are both absolutely continuous with respect to Π and given by

$$\begin{aligned}
\frac{d\Pi^m}{d\Pi}(u) &= \frac{1}{Z}\exp(-\Phi(u)), \quad Z = \int_X \exp(-\Phi(u))d\Pi(u) \quad \text{and} \\
\frac{d\Pi^m_N}{d\Pi}(u) &= \frac{1}{Z_N}\exp(-\Phi_N(u)), \quad Z_N = \int_X \exp(-\Phi_N(u))d\Pi(u).
\end{aligned} \tag{2.2}$$

The measure Π_N can arise e.g. when approximating the forward map A in (1.6). It is important to know whether closeness of the forward map and its approximation imply closeness of the posterior measure.

Assumption 2.12. Let $X' \subset X$ and assume that $\Phi \in C(X'; \mathbb{R})$. Assume further that there exists functions $M_i : \mathbb{R}^+ \to \mathbb{R}^+$, $i = 1, 2$ that are independent of N, non-decreasing and M_2 being strictly positive, such that for all $u \in X'$,

$$\begin{gathered}
\Phi(u) \geq -M_1(\|u\|_X), \quad \Phi_N(u) \geq -M_1(\|u\|_X) \quad \text{and} \\
|\Phi(u) - \Phi_N(u)| \leq M_2(\|u\|_X)\psi(N)
\end{gathered}$$

where $\psi(N) \to 0$ as $N \to \infty$.

The following theorems are similar to the ones in the previous section but they estimate changes in the posterior caused by changes in the potential Φ rather than data m.

Theorem 2.13. *Let Assumption 2.12 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(A \cap X') > 0$ *for some bounded set* A *in* X*. We also assume that*

$$\exp(M_1(\|u\|_X)) \in L^1_\Pi(X; \mathbb{R}).$$

Then Z *and* Z_N *defined in 2.2 are positive and finite, and the probability measures* Π^m *and* Π^m_N *are well defined. Furthermore, for sufficiently large* N*,* Z_N *is bounded below by a positive constant independent of* N*.*

Proof. The finiteness of Z and Z_N follows from the lower bounds on Φ and Φ_N given in Assumption 2.12 combined with the integrability condition assumed in the theorem. Since $u \sim \Pi$ satisfies $u \in X'$ a.s. we have

$$Z = \int_{X'} \exp(-\Phi(u)) d\Pi(u)$$

Note that $A' = A \cap X'$ is bounded in X and hence $\sup_{u \in A'} \|u\|_X = R_1 < \infty$. Since $\Phi : X' \to \mathbb{R}$ is continuous it is finite in every point of A'. Using the Assumption 2.12 for large enough N we can write

$$\sup_{u \in A'} |\Phi(u) - \Phi_N(u)| \leq R_2 < \infty.$$

This implies

$$\sup_{u \in A'} \Phi(u) = 2R_2 < \infty \quad \text{and} \quad \sup_{u \in A'} \Phi_N(u) = 2R_2 < \infty.$$

Hence

$$Z \geq \int_{A'} \exp(-2R_2) d\Pi(u) = \exp(-2R_2)\Pi(A') > 0.$$

We get the same lower bound for Z_N and note that it is independent of N as required. □

Theorem 2.14. *Let Assumption 2.12 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(A \cap X') > 0$ *for some bounded set* A *in* X*. We assume furthermore that*

$$\exp(M_1(\|u\|_X))\Big(1 + M_2(\|u\|_X)^2\Big) \in L^1_\Pi(X; \mathbb{R}).$$

Then there exists $c > 0$ *such that*

$$d_{Hell}(\Pi^m, \Pi^m_N) \leq c\psi(N)$$

for all sufficiently large N*.*

Proof. Let N be sufficiently large so that by Theorem 2.14 $Z > 0$ and $Z_N > 0$ with positive lower bounds independent of N. Using the local Lipschitz property of exponential, Assumption 2.12 and the fact that $M_2(\|u\|_X) \leq 1 + M_2(\|u\|_X)^2$ we can write

$$\begin{aligned}
|Z - Z_N| &\leq \int_{X'} |\exp(-\Phi(u)) - \exp(-\Phi_N(u))| d\Pi(u) \\
&\leq \int_{X'} \exp(M_1(\|u\|_X)) |\Phi(u) - \Phi_N(u)| d\Pi(u) \\
&\leq \psi(N) \int_{X'} \exp(M_1(\|u\|_X)) M_2(\|u\|_X) d\Pi(u) \\
&\leq \psi(N) \int_{X'} \exp(M_1(\|u\|_X))(1 + M_2(\|u\|_X)^2) d\Pi(u) \\
&\leq C\psi(N),
\end{aligned}$$

where C is a constant that does not depend on u or N. As in the proof of Theorem 2.9 we can write

$$\Big(d_{Hell}(\Pi^m, \Pi^m_N)\Big)^2 = I_1 + I_2,$$

where

$$\begin{aligned}
I_1 &= \frac{1}{Z} \int_{X'} \Big(\exp\big(-\frac{1}{2}\Phi(u)\big) - \exp\big(-\frac{1}{2}\Phi_N(u)\big)\Big)^2 d\Pi(u) \quad \text{and} \\
I_2 &= \big(Z^{-\frac{1}{2}} - Z_N^{-\frac{1}{2}}\big)^2 \int_{X'} \exp\big(\Phi_N(u)\big) d\Pi(u).
\end{aligned}$$

Using similar arguments as above we see that

$$\begin{aligned}
I_1 &\leq \frac{1}{4Z} \int \exp(M_1(\|u\|_X)) |\Phi(u) - \Phi_N(u)|^2 d\Pi(u) \\
&\leq \frac{\psi(N)^2}{Z} \int \exp(M_1(\|u\|_X)) M_2(\|u\|_X)^2 d\Pi(u) \\
&\leq C\psi(N)^2.
\end{aligned}$$

We also notice that

$$\int_{X'} \exp\big(\Phi_N(u)\big) d\Pi(u) \leq \int_{X'} \exp\big(M_1(\|u\|_X)\big) d\Pi(u) < \infty$$

and the upper bound is independent of N. Hence

$$I_2 \leq \frac{c\big(Z - Z_N\big)^2}{\min\big(Z^3, Z_N^3\big)} \leq C\psi(N)^2,$$

which concludes the proof. □

2.4 Infinite dimensional Gaussian measure

We start by introducing infinite dimensional Gaussian random variables and some of their key properties. For more details see e.g. [6, Section 3] or [3, Section 2], and if you feel brave [2].

Let X be a separable Banach space and denote by X^* its dual space of linear functionals on X. We define the characteristic function of a probability distribution μ on a separable Banach Space X as

$$\varphi_\mu(\psi) = \mathbb{E}\exp(i\psi(x)),$$

for $\psi \in X^*$.

Theorem 2.15. *If μ and ν are two probability measures on a separable Banach space X and if $\varphi_\mu(\psi) = \varphi_\nu(\psi)$, for all $\psi \in X^*$, then $\mu = \nu$.*

A function $\theta \in X$ is called the mean of μ if $\psi(\theta) = \int_X \psi(x)d\mu(x)$ for all $\psi \in X^*$. A linear operator $\Sigma : X^* \to X$ is called the covariance operator if $\psi(\Sigma\phi) = \int_X \psi(x - \theta)\phi(x - \theta)d\mu(x)$ for all $\psi, \phi \in X^*$. If we assume that $X = \mathcal{H}$ is a Hilbert space then $\theta = \mathbb{E}(x)$ and the covariance operator is characterised by identity $\mathbb{E}\big(\langle\phi, (x - \theta)\rangle\langle\psi, (x - \theta)\rangle\big) = \langle\Sigma\phi, \psi\rangle$.

A measure μ on $(X, \mathcal{B}(X))$ is Gaussian if, for any $\psi \in X^*$, $\psi(x) \sim \mathcal{N}(\theta_\psi, \sigma_\psi^2)$ for some $\theta_\psi \in \mathbb{R}$ and $\sigma_\psi \in \mathbb{R}$. We allow $\sigma_\psi = 0$, so that the measure may be a Dirac mass at θ_ψ. Note that it is expected that $\theta_\psi = \psi(\theta)$ and $\sigma_\psi^2 = \psi(\Sigma\psi)$ for all $\psi \in X^*$.

Theorem 2.16. *A Gaussian measure μ on $(X, \mathcal{B}(X))$ has a mean θ and covariance operator Σ. The characteristic function of the measure is*

$$\varphi_\mu(\psi) = \exp\left(i\psi(\theta) - \frac{1}{2}\psi(\Sigma\psi)\right).$$

Using the above Theorem and Theorem 2.15 we see that the mean and covariance completely characterise the Gaussian measure and hence we can simply write $\mathcal{N}(\theta, \Sigma)$.

Definition 2.17. *Let $\{\phi_i\}_{i=1}^\infty$ denote an orthonormal basis for a separable Hilbert space $\mathcal{H}$. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is trace-class if*

$$Tr(A) = \sum_{i=1}^{\infty}\langle A\phi_i, \phi_i\rangle < \infty.$$

The sum is independent of the choice of basis. The operator A is Hilbert–Schmidt if

$$Tr(A^*A) = \sum_{i=1}^{\infty} \|A\phi_i\|_{\mathcal{H}}^2 < \infty.$$

We can construct random draws from a Gaussian measure on a Hilbert space $\mathcal{H}$ using Karhunen–Loève expansion.

Theorem 2.18. *Let Σ be a self-adjoint, positive semi-definite, trace class operator in a Hilbert space $\mathcal{H}$, and let $\theta \in \mathcal{H}$. Let $\{\phi_k, \gamma_k\}$ be an orthonormal set of eigenvectors and eigenvalues for Σ ordered so that $\gamma_1 \geq \gamma_2 \geq \cdots$. Take $\{\xi_k\}_{k=1}^{\infty}$ to be an i.i.d. sequence with $\xi_1 \sim \mathcal{N}(0,1)$. Then the random variable $x \in \mathcal{H}$ given by the Karhunen–Loève expansion*

$$x = \theta + \sum_{k=1}^{\infty} \sqrt{\gamma_k}\xi_k\phi_k \tag{2.3}$$

is distributed according to $\mu = \mathcal{N}(\theta, \Sigma)$.

The proof is left as an exercise.

Example 2.19. A random variable η is said to be white Gaussian noise on $L^2(\mathbb{T}^d)$ if $\mathbb{E}(\eta) = 0$ and $\mathbb{E}\big(\langle \eta, \phi\rangle\langle \eta, \psi\rangle\big) = \langle \phi, \psi\rangle$, in which case we denote $\eta \sim \mathcal{N}(0, I)$. Note that $I : L^2(\mathbb{T}^d) \to L^2(\mathbb{T}^d)$ is not a trace class operator in $L^2(\mathbb{T}^d)$, and hence white noise does not take values in $L^2(\mathbb{T}^d)$. Let $e_{\vec{\ell}} \in L^2(\mathbb{T}^d)$, $\vec{\ell} = (\ell_1, \ell_2, \dots, \ell_d) \in \mathbb{Z}^d$ be an orthonormal basis of $L^2(\mathbb{T}^d)$ consisting of eigenfunctions of Laplacian, numbered so that $-\Delta e_{\vec{\ell}} = |\vec{\ell}|^2 e_{\vec{\ell}}$. Such functions $e_{\vec{\ell}}(x)$ can be chosen to be normalised products of the sine and cosine functions $\sin(\ell_j x_j)$ and $\cos(\ell_j x_j)$ that form the standard Fourier basis of $L^2(\mathbb{T}^d)$. The Fourier coefficients of η with respect to this basis are independent, normally distributed $\mathbb{R}$-valued random variables with variance one, that is, $\langle \eta, e_{\vec{\ell}}\rangle \sim N(0,1)$. Then

$$\mathbb{E}\|\eta\|^2_{L^2(\mathbb{T}^d)} = \sum_{\vec{\ell}\in\mathbb{Z}^d} \mathbb{E}|\langle \eta, e_{\vec{\ell}}\rangle|^2 = \sum_{\vec{\ell}\in\mathbb{Z}^d} 1 = \infty.$$

This implies that realisations of η are in $L^2(\mathbb{T}^d)$ with probability zero. However, when $s > d/2$

$$\mathbb{E}\|\eta\|^2_{H^{-s}(\mathbb{T}^d)} = \sum_{\vec{\ell}\in\mathbb{Z}^d} (1 + |\vec{\ell}|^2)^{-s} \mathbb{E}|\langle \eta, e_{\vec{\ell}}\rangle|^2 < \infty \tag{2.4}$$

and hence η takes values in $H^{-s}(\mathbb{T}^d)$ a.s. For more details about Sobolev spaces see Appendix A.

The above result can be generalised to show that if $x \sim \mathcal{N}(0, \Sigma)$ and the eigenvalues of Σ satisfy $\gamma_j \asymp j^{-\frac{2s}{d}}$ (e.g. $\Sigma = (I - \Delta)^{-s}$) then, for $t < s - d/2$, we have $x \in H^t$ a.s. We can also generalise the results for more general domains than the torus or $\mathbb{R}^d$ (or a closed manifold) using Hilbert scales. These spaces do not, in general, coincide with Sobolev spaces, because of the effect of the boundary conditions.

The covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ of a Gaussian on $\mathcal{H}$ is a compact operator and its inverse is densely defined unbounded operator on $\mathcal{H}$. We call this inverse precision operator. Bot the covariance and the precision operator are self-adjoint on appropriate domains and the fractional powers of them can be defined via spectral theorem.

Given a Gaussian measure μ on a separable Banach space X, we define the Cameron–Martin space $V_\mu \subset X$ of μ to be the intersection of all linear spaces of full measure. The main importance of the Cameron-Martin space is that it characterises exactly the directions in X in which a centred Gaussian measure can be shifted to obtain an equivalent Gaussian measure. When $\dim(X) = \infty$ the measure of the Cameron–Martin space is zero, that is, $\mu(V_\mu) = 0$. Compare this to the case of finite dimensional Lebesgue measure which is invariant under translations in any direction. This is a striking illustration of the fact that measures in infinite-dimensional spaces have a strong tendency of being mutually singular.

Lemma 2.20. *For a Gaussian measure on Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ the Cameron–Martin space V_μ consists of the image of $\mathcal{H}$ under $\Sigma^{\frac{1}{2}}$ and the Cameron–Martin norm is given by $\|h\|_\mu^2 = \|\Sigma^{-\frac{1}{2}}h\|_{\mathcal{H}}^2$.*

Theorem 2.21. *Let $\mu = \mathcal{N}(0,\Sigma)$ be a Gaussian measure on a separable Banach space X. The Cameron–Martin space V_μ of μ can be endowed with Hilbert space structure and V_μ is compactly embedded in all separable spaces X' such that $\mu(X') = 1$.*

Theorem 2.22 (Special case of the Cameron-Martin theorem)**.** *Let $\mu = \mathcal{N}(0,\Sigma)$ be a Gaussian measure on a separable Banach space X. Denote by μ_h the translation of μ by h, $\mu_h = \mu(\cdot - h)$. If $h \in V_\mu$ then μ_h is absolutely continuous with respect to μ and*

$$\frac{d\mu_h}{d\mu}(x) = \exp\Big(-\frac{1}{2}\|h\|_{V_\mu}^2 + \langle h, x\rangle_{V_\mu}\Big)$$

$x \in X$, μ-a.s. If $h \notin V_\mu$, then μ and μ_h are mutually singular.

Example 2.23. Consider two Gaussian measures μ_i, $i = 1,2$, on $\mathcal{H} = L^2((0,1))$ both with precision operator (the densely defined inverse covariance operator $\Sigma^{-1} = \mathcal{L}$) $\mathcal{L} = -d^2/dx^2$, the domain of $\mathcal{L}$ being $H_0^1((0,1)) \cap H^2((0,1))$. We assume that $\mu_1 \sim \mathcal{N}(\theta,\Sigma)$ and $\mu_2 \sim \mathcal{N}(0,\Sigma)$. Then $V_\mu = \mathrm{Im}(\Sigma^{1/2}) = H_0^1((0,1))$. Hence the measures are equivalent if and only if $\theta \in V_\mu$. If this is satisfied then the Radon–Nikodym derivative between the two measures is given by

$$\frac{d\mu_1}{d\mu_2}(x) = \exp\Big(\langle\theta, x\rangle_{H_0^1} - \frac{1}{2}\|\theta\|_{H_0^1}^2\Big).$$

Example 2.24. Let us return to the Example 1.2 where we wanted to recover the initial condition u of the heat equation from a noisy observation m of the solution at some time $T > 0$. We have the observation model

$$m = Au + \eta,$$

where $A = e^{\Delta} : L^2(\mathbb{T}^d) \to L^2(\mathbb{T}^d)$.

We place a Gaussian prior $\Pi = \mathcal{N}(\theta,\Sigma)$ on u, with $\Sigma = (I-\Delta)^{-\alpha}$, for some $\alpha > \frac{d}{2}$. Then $\Pi(H^s) = 1$ for all $s < \alpha - \frac{d}{2}$ and especially $\Pi(X) = \Pi(L^2) = 1$. For the noise we assume that $\eta \sim P_0 = \mathcal{N}(0, \delta^2 I)$, that is, η is white Gaussian noise with noise amplitude δ. This measure satisfies $P_0(H^s) = 1$ for $s < -\frac{d}{2}$ and we can choose

$Y = H^{s'}$ for some $s' < -\frac{d}{2}$, $m \in Y$ a.s. The Cameron–Martin space of white Gaussian noise is L^2.

We then have $Au \in V_{P_0}$ and $m \,|\, u \sim P_u = \mathcal{N}(Au, \delta^2 I)$, where $P_u \ll P_0$, with

$$\frac{dP_u}{dP_0}(m) = \exp(-\Phi(u;m))$$

and

$$\Phi(u;m) = \frac{1}{2}\|Au\|_{L^2}^2 - \langle m, Au\rangle_{L^2}.$$

In the following we repeatedly use the fact that $\Delta^\gamma e^{\kappa\Delta}$, $\kappa > 0$, is a bounded linear operator from H^a to H^b for any $a, b, \gamma \in \mathbb{R}$. Note that $\nu_0(L^2 \times H^{s'}) = 1$, where we have denoted $\nu_0(du, dm) = \Pi(du)P_0(dm)$. Using the boundedness of $\Delta^\gamma e^{\kappa\Delta}$ we can show that

$$\Phi : L^2 \times H^{s'} \to \mathbb{R}$$

is bounded, and hence by Lemma 2.2 ν_0-measurable.

We can then use Theorem 2.3 to conclude that the posterior is given by Π^m, where

$$\begin{aligned} \frac{d\Pi^m}{d\Pi}(u) &= \frac{1}{Z}\exp(-\Phi(u;m)) \\ Z &= \int_{L^2} \exp(-\Phi(u;m))d\Pi(u), \end{aligned}$$

provided that $Z = Z(m) > 0$ P_0-a.s. As stated before $m \in H^s$ for any $s < -\frac{d}{2}$ P_0-a.s. and hence $m = \Delta^{-\frac{s'}{2}} w$ with some $w \in L^2$ and $s' < -\frac{d}{2}$. We can then write

$$\Phi(u;m) = \frac{1}{2}\|Au\|_{L^2}^2 - \langle e^{\frac{1}{2}\Delta}\Delta^{-\frac{s'}{2}} w, e^{\frac{1}{2}\Delta} u\rangle_{L^2}.$$

Using the above formulation together with the boundedness of $\Delta^\gamma e^{\kappa\Delta}$, $\kappa > 0$, we get

$$\Phi(u;m) \le C(\|w\|_{L^2}^2 + \|u\|_{L^2}^2),$$

where $\|w\|_{L^2} < \infty$ P_0-a.s. Thus

$$Z \ge \int_{\|u\|_{L^2} \le 1} \exp\big(-C(\|w\|_{L^2}^2 + 1)\big) d\Pi(u)$$

and since all balls have positive Gaussian measure on a separable Banach spaces we see that $Z > 0$.

2.5 MAP estimators and Tikhonov regularisation

In this section we assume that the prior Π is Gaussian. We show that MAP estimators (point of maximal probability) coincide with the minimisers of Tikhonov regularised least squares functions with regularisation term being given by the Cameron-Martin norm of the Gaussian prior.

The classical deterministic way to solve inverse problem is to try to minimise the potential Φ with some regularisation. If we had finite data and Gaussian observational noise $\eta \sim N(0, \Gamma)$ we can write

$$\Phi(u; m) = \frac{1}{2} \left\| \Gamma^{-1/2}(m - Au) \right\|^2.$$

Thus Φ is covariance weighted data misfit least square function.

We assume that Π is a Gaussian probability measure on a separable Banach space $(X, \|\cdot\|_X)$ and $\Pi(X) = 1$. We denote the Cameron-Martin space of Π by $(V_\Pi, \|\cdot\|_{V_\Pi})$. In this section we want to show that maximising Π^m is equivalent to minimising

$$I(u) = \begin{cases} \Phi(u; m) + \frac{1}{2}\|u\|_{V_\Pi}^2 & \text{if } u \in V_\Pi, \text{ and} \\ \infty & \text{else.} \end{cases} \tag{2.5}$$

The realisation m of the data does not play role in this section and we will write $\Phi(u; m) = \Phi(u)$.

We note that the properties of Φ we assume below are typically determined by the forward operator, which maps the unknown function u to the data m. Probability theory does not play a direct role in verifying these properties of Φ. Probability becomes relevant when choosing the prior measure Π so that it charges the Banach space X, on which the desired properties of Φ hold, with full measure.

Assumption 2.25. The function $\Phi : X \to \mathbb{R}$ satisfies the following conditions:

1) For every $\varepsilon > 0$ there is an $R = R(\varepsilon) \in \mathbb{R}$, such that for all $u \in X$,

$$\Phi(u) \geq R - \varepsilon \|u\|_X^2.$$

2) Φ is locally bounded above, that is, for every $r > 0$ there exists $K = K(r) > 0$ such that, for all $u \in X$ with $\|u\|_X < r$, we have

$$\Phi(u) \leq K.$$

3) Φ is locally Lipschitz continuous i.e. for every $r > 0$ there exists $L = L(r) > 0$ such that, for all $u_1, u_2 \in X$ with $\|u_1\|_X, \|u_2\|_X < r$, we have

$$|\Phi(u_1) - \Phi(u_2)| \leq L \|u_1 - u_2\|_X.$$

In finite dimensions there is an obvious notion of most likely points for measures which have a continuous density with respect to Lebesgue measure: the points at which the Lebesgue density is maximised. Unfortunately we can not translate this idea to infinite dimensions. To fix this we will restate the idea in a way that will work also in infinite dimensional settings. Fix a small radius $\delta > 0$ and identify centres of balls of radius δ which have maximal probability. Letting $\delta \to 0$ then recovers the maximums when there is continuous Lebesgue density. We will use this small ball approach in infinite dimensional settings.

Let $z \in V_\Pi$ and $B_\delta(z) \subset X$ be the open ball centred at $z \in X$ with radius δ in X. Let

$$J_\delta^m(z) = \Pi^m(B_\delta(z))$$

be the mass of the ball $B_\delta(z)$ under the posterior measure Π^m. Similarly we define

$$J_\delta(z) = \Pi(B_\delta(z))$$

to be the mass of the ball $B_\delta(z)$ under the Gaussian prior. We note that all balls in a separable Banach space have positive Gaussian measure. Thus $J_\delta(z)$ is finite and positive for any $z \in V_\Pi$. By the above assumptions on Φ and the Fernique Theorem 2.11 the same is true for $J_\delta^m(z)$. We will next prove that the probability is maximised where I is minimised.

Theorem 2.26. *Let Assumption 2.25 hold and assume that* $\Pi(X) = 1$*. Then, for any* $z_1, z_2 \in V_\Pi$,

$$\lim_{\delta \to 0} \frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} = \exp\big(I(z_1) - I(z_1)\big),$$

where the function I *is defined by* (2.5).

Before moving to prove the above theorem we state a result about the small ball probabilities under Gaussian measure

Theorem 2.27. *Let* $z \in V_\Pi$ *and* $B_\delta(z) \subset X$ *be the open ball centred at* $z \in X$ *with radius* δ *in* X*. The ratio of small ball probabilities under Gaussian measure* Π *satisfies*

$$\lim_{\delta \to 0} \frac{\Pi(B_\delta(z_1))}{\Pi(B_\delta(z_2))} = \exp\left(\frac{1}{2}\|z_2\|_{V_\Pi}^2 - \frac{1}{2}\|z_1\|_{V_\Pi}^2\right).$$

Proof of theorem 2.26. The ratio is finite and positive since $J_\delta^m(z)$ is finite and positive far any $z \in V_\Pi$. The estimate given in Theorem 2.27 transfers the question about probability into statement concerning the Cameron-Martin norm of Π. Note that if $u \sim \Pi$ then its realisation is in V_Π only with probability zero and hence $\|u\|_{V_\Pi} = \infty$ almost surely.

We can write

$$\begin{aligned}\frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} &= \frac{\int_{B_\delta(z_1)} \exp(-\Phi(u))d\Pi(u)}{\int_{B_\delta(z_2)} \exp(-\Phi(v))d\Pi(v)} \\ &= \frac{\int_{B_\delta(z_1)} \exp(-\Phi(u)+\Phi(z_1))\exp(-\Phi(z_1))d\Pi(u)}{\int_{B_\delta(z_2)} \exp(-\Phi(v)+\Phi(z_2))\exp(-\Phi(z_2))d\Pi(v)}.\end{aligned}$$

By Assumption 2.25 there exists $L = L(r)$ such that

$$-L\|u_1 - u_2\|_X \le \Phi(u_1) - \Phi(u_2) \le L\|u_1 - u_2\|_X$$

for all $u_1, u_2 \in X$ with $\max\{\|u_1\|_X, \|u_2\|_X\} < r$. We can then write

$$\begin{aligned}\frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} &\le e^{2\delta L} \frac{\int_{B_\delta(z_1)} \exp(-\Phi(z_1)) d\Pi(u)}{\int_{B_\delta(z_2)} \exp(-\Phi(z_2)) d\Pi(v)} \\ &\le e^{2\delta L} e^{-\Phi(z_1)+\Phi(z_2)} \frac{\Pi(B_\delta(z_1))}{\Pi(B_\delta(z_2))},\end{aligned}$$

Using Theorem 2.27 we get

$$\frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} \le r_1(\delta) e^{2\delta L} e^{-I(z_1)+I(z_2)}$$

where $r_1(\delta) \to 1$ as $\delta \to 0$. Thus

$$\limsup_{\delta \to 0} \frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} \le e^{-I(z_1)+I(z_2)}.$$

We can deduce in the same way that

$$\frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} \ge r_2(\delta) e^{-2\delta L} e^{-I(z_1)+I(z_2)}$$

with $r_2(\delta) \to 1$ as $\delta \to 0$ and furthermore

$$\liminf_{\delta \to 0} \frac{J_\delta^m(z_1)}{J_\delta^m(z_2)} \ge e^{-I(z_1)+I(z_2)},$$

which concludes the proof. □

We will next show that the minimisation problem for I is well-defined when Assumption 2.25 holds.

Definition 2.28. *Let E be a Hilbert space. The function $I : E \to \mathbb{R}$ is weakly lower semicontinuous if*

$$\liminf_{j \to \infty} I(u_j) \ge I(u)$$

whenever $u_j \rightharpoonup u$ in E. The function $I : E \to \mathbb{R}$ is weakly continuous if

$$\lim_{j \to \infty} I(u_j) = I(u)$$

whenever $u_j \rightharpoonup u$ in E.

Lemma 2.29. *Let $(E, \langle \cdot, \cdot \rangle_E)$ be a Hilbert space with induced norm $\| \cdot \|_E$. Then the quadratic form $J(u) = \frac{1}{2}\|u\|_E$ is weakly lower semicontinuous.*

Proof. We can write

$$\begin{aligned}
J(u_j) - J(u) &= \frac{1}{2}\|u_j\|_E^2 - \frac{1}{2}\|u\|_E^2 \\
&= \frac{1}{2}\langle u_j - u, u_j + u\rangle_E \\
&= \frac{1}{2}\langle u_j - u, 2u\rangle_E + \frac{1}{2}\|u_j - u\|_E^2 \\
&\geq \frac{1}{2}\langle u_j - u, 2u\rangle_E \to 0,
\end{aligned}$$

when $u_j \rightharpoonup u$ in E. □

Theorem 2.30. *Suppose that Assumption 2.25 holds and let E be a Hilbert space compactly embedded in X. Then there exists $\overline{u} \in E$ such that*

$$I(\overline{u}) = \overline{I} := \inf\{I(u) \in E\}.$$

Furthermore if $\{u_j\}$ is a minimising sequence satisfying $I(u_j) \to I(\overline{u})$ then there exists a subsequence $\{u_{j'}\}$ that converges strongly to $\overline{u}$ in E.

Proof. Compactness of $E \subset X$ implies that $\|u\|_X \leq C\|u\|_E$. Hence by Assumption 2.25 1) it follows that for any $\varepsilon > 0$ there is $R(\varepsilon) \in \mathbb{R}$ such that

$$I(u) \geq \Big(\frac{1}{2} - \varepsilon C\Big)\|u\|_E^2 + R(\varepsilon).$$

We can choose ε small enough so that

$$I(u) \geq \frac{1}{4}\|u\|_E^2 + R \tag{2.6}$$

for all $u \in E$ with some $R \in \mathbb{R}$.

Let u_j be minimising sequence satisfying $I(u_j) \to I(\overline{u})$ as $j \to \infty$. For any $\delta > 0$ there is $N = N(\delta)$, such that for all $j \geq N$

$$\overline{I} \leq I(u_j) \leq \overline{I} + \delta. \tag{2.7}$$

We can then use (2.6) to conclude that $\{u_j\}$ is bounded in E. We assumed that E is a Hilbert space so there exists $\overline{u} \in E$ such that $u_j \rightharpoonup \overline{u}$ in E. Since E is compactly embedded in X we can deduce that $u_j \to \overline{u}$ strongly in X. By the Assumption 2.25 3) the potential Φ is Lipschitz continuous and hence $\Phi(u_j) \to \Phi(\overline{u})$. Thus Φ is weakly continuous on E. Using Lemma 2.29 we see that $I(u) = J(u) + \Phi(u)$ is weakly lower semicontinuous on E. Using (2.7) we can then conclude that, for any $\delta > 0$,

$$\overline{I} \leq I(\overline{u}) \leq \overline{I} + \delta.$$

Since δ can be chosen arbitrarily small the first result follows.

Next we study a subsequence of u_j. For large enough n, ℓ we can write

$$\begin{aligned}\frac{1}{4}\|u_n - u_\ell\|_E^2 &= \frac{1}{2}\|u_n\|_E^2 + \frac{1}{2}\|u_\ell\|_E^2 - \frac{1}{4}\|u_n + u_\ell\|_E^2 \\ &= I(u_n) + I(u_\ell) - 2I\Big(\frac{1}{2}(u_n + u_\ell)\Big) - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\Big(\frac{1}{2}(u_n + u_\ell)\Big) \\ &\le 2(\overline{I} + \delta) - 2\overline{I} - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\Big(\frac{1}{2}(u_n + u_\ell)\Big) \\ &\le 2\delta - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\Big(\frac{1}{2}(u_n + u_\ell)\Big).\end{aligned}$$

The subsequences u_n, u_ℓ and $\frac{1}{2}(u_n + u_\ell)$ converge strongly to $\overline{u} \in X$. Since Φ is continuous we see that for large enough n, ℓ

$$\frac{1}{4}\|u_n - u_\ell\|_E^2 \le 3\delta.$$

We have shown that the subsequence is Cauchy in E which completes the proof. □

Note that by Theorem 2.21 the Cameron–Martin space V_Π is a Hilbert space that is compactly embedded in X and hence we can find a minimiser in V_Π.

3 Behaviour of the posterior distribution

3.1 Posterior consistency

We return to the discrete setting to study the posterior distribution in more detail. We assume Gaussian noise and prior and analyse how the posterior distribution behaves when the noise tends to zero.

Assumption 3.1. We assume that

- $u \sim \mathcal{N}(0, \Sigma_u)$, where Σ_u is symmetric and positive definite, and $u \perp \eta$.
- $\eta = \delta\eta_0$, with $\eta_0 \sim \mathcal{N}(0, \Gamma_0)$,where Γ_0 is symmetric and positive definite.
- $A \in \mathbb{R}^{k\times k}$ is invertible.
- $m^\dagger = Au^\dagger + \delta^2\eta_0^\dagger$, with a fixed $u^\dagger, \eta^\dagger \in \mathbb{R}^k$.

Theorem 3.2. *Let Assumption 3.1 hold. Then, for any sequence $C(\delta) \to \infty$ as $\delta \to 0$,*

$$\Pi^{m^\dagger}\big(\|u - u^\dagger\|^2 > C(\delta)\delta^2\big) \to 0.$$

Note that we can set $C(\delta) = \frac{\varepsilon}{\delta^2}$ to obtain

$$\Pi^{m^\dagger}\big(\|u - u^\dagger\| > \varepsilon\big) \to 0,$$

as the noise tends to zero. Hence Theorem 3.2 implies that u converges to $u^\dagger$ in probability.

Proof. Since $u \sim \mathcal{N}(0, \Sigma_u)$ and $\eta \sim \mathcal{N}(0, \delta^2\Gamma_0)$ we know that the posterior is also Gaussian and $u \,|\, m^\dagger \sim \mathcal{N}(\theta, \Sigma)$, where $\theta = (A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})^{-1}A^\top\Gamma_0^{-1}m^\dagger$ and $\Sigma = \delta^2(A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})^{-1}$. Denote by v the centred Gaussian random variable $v = u \,|\, m - \theta \sim \mathcal{N}(0, \Sigma)$. Let $\mathbb{E}$ denote the expectation with respect to the posterior distribution when we are given a measurement $m^\dagger$. We can then write

$$\begin{aligned}\mathbb{E}(\|u - u^\dagger\|^2) &= \mathbb{E}(\|\theta - u^\dagger + v\|^2)\\ &= \mathbb{E}(\|\theta - u^\dagger\|^2) + \mathbb{E}(\|v\|^2)\\ &= \|\theta - u^\dagger\|^2 + \mathrm{Tr}(\Sigma).\end{aligned}$$

We start by approximating the first term on the right hand side. Let $\{\lambda_j^2, \varphi_j\}$, $j = 1, \cdots, k$, be the eigenvalues and orthogonal eigenvectors of Σ_u, ordered so that $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_k \geq 0$. We will use the finite dimensional version of Karhunen–Loève expansion and write the k-dimensional random variable $x \sim \mathcal{N}(0, \Sigma_u)$ in form

$$\sum_{j=1}^{k} \lambda_j \xi_j \varphi_j,$$

where $\{\xi_j\}_{j=1}^k$ is a collection of independent $\mathcal{N}(0, 1)$ random variables.

From the definition of the posterior mean we see that

$$\begin{aligned}(A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})\theta &= A^\top\Gamma_0^{-1}m^\dagger\\ &= A^\top\Gamma_0^{-1}(Au^\dagger + \delta\eta_0^\dagger).\end{aligned}$$

Subtracting $(A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})u^\dagger$ from both sides and denoting $e = \theta - u^\dagger$ we can write

$$(A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})e = \delta A^\top\Gamma_0^{-1}\eta_0^\dagger - \delta^2\Sigma_u^{-1}u^\dagger.$$

Taking the inner product of both sides with e we obtain

$$\langle e, A^\top\Gamma_0^{-1}Ae\rangle + \delta^2\langle e, \Sigma_u^{-1}e\rangle = \delta\langle e, A^\top\Gamma_0^{-1}\eta_0^\dagger\rangle - \delta^2\langle e, \Sigma_u^{-1}u^\dagger\rangle,$$

which simplifies to

$$\|Ae\|_{\Gamma_0}^2 + \delta^2\|e\|_{\Sigma_u}^2 = \delta\langle e, A^\top\Gamma_0^{-1}\eta_0^\dagger\rangle - \delta^2\langle e, \Sigma_u^{-1}u^\dagger\rangle.$$

Since the matrix A was assumed to be invertible $\|A\cdot\|_{\Gamma_0}$ defines a norm in $\mathbb{R}^k$. All norms in $\mathbb{R}^k$ are equivalent and hence there exists $\alpha = \alpha(A, \Gamma_0)$ such that $\|Ae\|_{\Gamma_0}^2 \geq \alpha\|e\|^2$. Since $\delta^2\|e\|_{\Sigma_u}^2 \geq 0$ the left hand side is larger or equal to $\alpha\|e\|^2$. To deal with the right hand side we denote

$$K = K(A, \Gamma_0, \Sigma_u, u^\dagger, \eta^\dagger) = 2\max(\|A^\top\Gamma_0^{-1}\eta_0^\dagger\|, \|\Sigma_u^{-1}u^\dagger\|).$$

Note that for a fixed realisation $m^\dagger$ the above K is a constant. Using Cauchy-Schwartz inequality we see that for $\delta < 1$

$$\begin{aligned}\delta\langle e, A^\top\Gamma_0^{-1}\eta_0^\dagger\rangle - \delta^2\langle e, \Sigma_u^{-1}u^\dagger\rangle &\leq \delta\|e\|\|A^\top\Gamma_0^{-1}\eta_0^\dagger\| + \delta^2\|e\|\|\Sigma_u^{-1}u^\dagger\|\\ &\leq \frac{\delta}{2}\|e\|K + \frac{\delta^2}{2}\|e\|K\\ &\leq \delta\|e\|K.\end{aligned}$$

Combining the above we get

$$\alpha\|e\|^2 \leq \delta K\|e\|,$$

which implies

$$\|e\| \leq \frac{\delta K}{\alpha}.$$

Next we move to approximate $\mathbb{E}\|v\|^2 = \mathrm{Tr}(\Sigma)$. Let us denote $b = \Sigma^{-1}a$. Since $\Sigma = \delta^2(A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})^{-1}$ we can write

$$\delta^2 b = (A^\top\Gamma_0^{-1}A + \delta^2\Sigma_u^{-1})a.$$

Taking the inner product of both sides with a we obtain

$$\delta^2\langle a, b\rangle = \|Aa\|_{\Gamma_0} + \delta^2\|a\|_{\Sigma_u}.$$

As before the right hand side is bounded below by $\alpha\|a\|^2$ and the left hand side is bounded above by $\delta^2\|a\|\|b\|$. Thus we get

$$\frac{\|a\|}{\|b\|} \leq \frac{\delta^2}{\alpha}$$

We finally see that

$$\|\Sigma\| = \sup_{b\in\mathbb{R}^k}\frac{\|\Sigma b\|}{\|b\|} = \sup_{b\in\mathbb{R}^k}\frac{\|a\|}{\|b\|} \leq \frac{\delta^2}{\alpha}.$$

Note that since Σ is a symmetric positive-definite matrix its eigenvalues are all positive real numbers that can be ordered $\gamma_1^2 \geq \gamma_2^2 \cdots \geq \gamma_k^2 \geq 0$, and $\gamma_1^2 = \|\Sigma\| \leq \frac{\delta^2}{\alpha}$. Hence we can conclude

$$\mathrm{Tr}(\Sigma) = \sum_{j=1}^{k}\gamma_j^2 \leq \frac{k\delta^2}{\alpha}$$

We now see that, for $\delta < 1$, we have

$$\mathbb{E}(\|u - u^\dagger\|^2) = \|e\|^2 + \mathrm{Tr}(\Sigma) \leq L\delta^2,$$

where $L = \frac{K^2+\alpha k}{\alpha^2}$ is a constant. Using the Markov inequality we conclude that, for any $C(\delta) \to \infty$ as $\delta \to 0$,

$$\Pi^m\big(\|u - u^\dagger\|^2 > C(\delta)\delta^2\big) \leq \frac{\mathbb{E}\|u - u^\dagger\|^2}{C(\delta)\delta^2} \leq \frac{L\delta^2}{C(\delta)\delta^2} \to 0$$

as $\delta \to 0$. □

3.2 Uncertainty quantification

In the previous sections we studied the problem of approximating an unknown u from a measurement $m = Au + \eta$. The solution to this Bayesian inverse problem is a posterior distribution. This distribution can be used to achieve a point estimator, however, the strength of Bayesian approach is that we can use posterior distribution to quantify how certain we are of the solution. We call a subset $\mathcal{C} \subset X$, such that

$$\Pi(u \in \mathcal{C} \,|\, m) = 1 - \alpha,$$

a $1 - \alpha$ level credible set for u. This approach seems attractive since attaining such credible sets is usually computationally cheap. In most applications we can not calculate the CM estimator explicitly but sampling from the posterior is possible. These samples can then be used to approximate the CM estimator, and furthermore to find sets with high posterior probability.

Another closely related statistical way of tackling inverse problems is the so called frequentist approach where we assume that the data is generated from a deterministic 'true' unknown $u^\dagger$, that is we have $m^\dagger = Au^\dagger + \eta$. We are then interested in how frequently the true unknown falls in some subsets. That is, a frequentist $1 - \alpha$ confidence region is defined as a subset $\mathcal{C}_\dagger = \mathcal{C}(m^\dagger) \in X$, such that

$$\mathbb{P}(u^\dagger \in \mathcal{C}_\dagger) = 1 - \alpha.$$

The objective meaning of the frequentist confidence regions is well understood but they are difficult to achieve when the parameter space is large as in most inverse problems. That is why we would like to know if the Bayesian credible sets have correct frequentist coverage.

We will study the measurement model

$$m^\dagger = Au^\dagger + \eta,$$

where $u^\dagger \in \mathbb{R}^d$ and $m, \eta \in \mathbb{R}^k$. If $\eta \sim P_0 = \mathcal{N}(0, I)$ then $m^\dagger \sim P_{u^\dagger} = \mathcal{N}(Au^\dagger, I)$.

Next we define Fisher information matrix which measures how much information the measurement $m^\dagger$ carries of the unknown $u^\dagger$.

Definition 3.3. *Fisher information matrix is defined as*

$$\mathcal{I}(u) = -\mathbb{E}\left(\frac{\partial^2}{\partial u_i \partial u_j} \log \rho(m - Au)\right)_{i,j=1}^{k}.$$

With our assumption $m^\dagger \sim \mathcal{N}(Au^\dagger, I)$ the Fisher information matrix is given by $\mathcal{I}(u^\dagger) = A^*A$.

We are interested in comparing the Bayesian and frequentist solutions. We will model the unknown u using a prior Π. To compare the Bayesian credible sets and the frequentist confidence regions we will need the following Bernstein–von Mises theorem which states that, on the small noise limit, the posterior distribution behaves like a normal distribution centred at an efficient estimator, such as the maximum likelihood estimator (MLE) $\widehat{u}_\delta$. We note that, as $\delta \to 0$, $\frac{1}{\delta}(\widehat{u}_\delta - u^\dagger) \to \mathcal{N}(0, \mathcal{I}(u^\dagger))$

Theorem 3.4 (Bernstein–von Mises)**.** *Let $m_\delta^\dagger \sim P_{u^\dagger} = \mathcal{N}(Au^\dagger, \delta^2 I)$. We assume that $u \sim \Pi$, where the prior has continuous density π at $u^\dagger$ with $\pi(u^\dagger) > 0$. Denote the associated posterior $\Pi_\delta^m = \Pi(\cdot \,|\, m_\delta^\dagger)$. Let μ_δ be distribution $\mathcal{N}\big(\widehat{u}_\delta, \delta^2 \mathcal{I}(u^\dagger)^{-1}\big)$. Then*

$$\|\Pi_\delta^m - \mu_\delta\|_{TV} = \int_X |\Pi_\delta^m(u) - \mu_\delta(u)| du \to 0 \quad a.s.$$

as $\delta \to 0$.

The above result states that the two distributions Π_δ^m and μ_δ look increasingly alike when $\delta \to 0$. This implies that for any subset A, we have $\Pi_\delta^m(A) - \mu_\delta(A) \to 0$, almost surely. As a consequence we have that, for any $1-\alpha$ level Bayesian credible set $\mathcal{C}_\delta$, $\mu_\delta(\mathcal{C}_\delta) \to 1-\alpha$. This is helpful in showing that the credible sets are frequentist confidence regions of level $1-\alpha$.

The BvM theorem can be stated in more general non-Gaussian settings assuming that the likelihood $\rho_u^\dagger(m)$ fulfils *the usual regularity assumptions*. For more details see e.g. the course notes of *Principles of Statistics*.

Next we want to show that credible sets of the form

$$\mathcal{C}_\delta = \Big\{ u : |\widehat{u}_\delta - u| \leq \delta R_\delta \Big\},$$

with R_δ chosen so that $\Pi_\delta^m(\mathcal{C}_\delta) = \Pi(\mathcal{C}_\delta \,|\, m_\delta^\dagger) = 1-\alpha$, are also frequentist confidence regions, i.e. for $m^\dagger \sim P_{u^\dagger}$ we have that $\mathbb{P}(u^\dagger \in \mathcal{C}_\delta) = 1-\alpha$, when $\delta \to 0$. We will start by shoving that if R_δ converges almost surely to its frequentist equivalent, the probability converge to $1-\alpha$. After that we show that R_δ indeed converges to this limit.

Definition 3.5. *Define the function F, for all $t \geq 0$, as*

$$F(t) = \mathbb{P}(|Z| \leq t) = \int_{-t}^{t} f(x) dx,$$

where $Z \sim \mathcal{N}(0, \mathcal{I}(u^\dagger)^{-1})$. Then $F : [0,\infty) \to [0,1)$ is an increasing, continuous one-to-one mapping and its well-defined functional inverse is also continuous and denoted by F^{-1}.

Lemma 3.6. *Under the above assumptions we have that $R_\delta \to F^1(1-\alpha)$ a.s., as $\delta \to 0$.*

Proof. Using the change of variables $x = \frac{1}{\delta}(u - \widehat{u}_\delta)$ we see that

$$F(R_\delta) = \int_{-R_\delta}^{R_\delta} f(x) dx = \int_{\widehat{u}_\delta - \delta R_\delta}^{\widehat{u}_\delta + \delta R_\delta} \mu_\delta(u) du = \mu_\delta(\mathcal{C}_\delta)$$

Hence, as $\delta \to 0$, we have $F(R_\delta) = \mu_\delta(\mathcal{C}_\delta) - \Pi_\delta^m(\mathcal{C}_\delta) + \Pi_\delta^m(\mathcal{C}_\delta) \to 1-\alpha$ a.s., since by the Bernstein–von Mises theorem, the first difference converges to 0 a.s. We conclude the proof by applying the continuous mapping theorem with F^{-1}. □

Theorem 3.7. *Let $\alpha \in (0,1)$. Then under the above assumptions we have that $P_{u^\dagger}(u^\dagger \in \mathcal{C}_\delta) \to 1-\alpha$.*

Proof. Given that $F^{-1}(1-\alpha) > 0$ we have that

$$\frac{F^{-1}(1-\alpha)}{R_\delta}\frac{1}{\delta}(\widehat{u}_\delta - u^\dagger) \to^d \mathcal{N}(0, \mathcal{I}(u^\dagger)^{-1}).$$

Hence we can write

$$\begin{aligned} P_{u^\dagger}(u^\dagger \in \mathcal{C}_\delta) &= P_{u^\dagger}\Big(|\widehat{u}_\delta - u^\dagger| \le \delta R_\delta\Big) \\ &= P_{u^\dagger}\bigg(\frac{F^{-1}(1-\alpha)}{R_\delta}\frac{1}{\delta}|\widehat{u}_\delta - u^\dagger| \le F^{-1}(1-\alpha)\bigg) \\ &\to \mathbb{P}\big(|Z| \le F^{-1}(1-\alpha)\big) = F\big(F^{-1}(1-\alpha)\big) = 1-\alpha, \end{aligned}$$

which concludes the proof. □

Note that we can replace the MLE $\widehat{u}_\delta$ by the CM estimator $\bar{u}_\delta$ in the above calculations.

Theorem 3.7 implies that on the small noise limit (or equivalently on the large data sample limit $n \to \infty$) the Bayesian credible sets have correct frequentist coverage. We note that the same is not true in infinite dimensional setting.

A Sobolev spaces

Sobolev spaces constitute one of the most relevant functional settings for the treatment of PDEs and boundary value problems. This appendix gives a short introduction to the topic. Sobolev spaces are covered properly on course *Analysis of Partial Differential Equations*. For more a more detailed treatment of Sobolev spaces and applications to PDEs see [5]. For a comprehensive study of Sobolev spaces see e.g. [1].

We start by introducing the notion of a weak derivatives that generalises the classical partial derivatives.

Definition A.1 (Test functions). *Let $\mathcal{O} \in \mathbb{R}^d$. We set*

$$C_0^\infty(\mathcal{O}) = \{\phi \in C^\infty(\mathcal{O}) : supp(\phi) \in V \subset \mathcal{O}\},$$

the smooth functions with compact support. This space is often referred as the space of test functions and denoted by $\mathcal{D}(\mathcal{O})$.

If $u \in C^1(\mathbb{R})$ then we can define $\frac{\partial u}{\partial x}$ by

$$\int \frac{\partial u}{\partial x}(x)\phi(x)dx = -\int u(x)\frac{\partial \phi}{\partial x}(x)dx,$$

for all $\phi \in \mathcal{D}(\mathbb{R})$. We notice that the right hand side is well-defined for all $u \in L^1_{loc}(\mathbb{R})$.

Definition A.2. *Let $\alpha = \alpha_1, \cdots, \alpha_d$ be a multi-index, $\alpha_i \in \mathbb{N}$, and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. A function $u \in L^1_{loc}(\mathcal{O})$ has a weak derivative $v = D^\alpha u \in L^1_{loc}(\mathcal{O})$ if*

$$\int_{\mathcal{O}} v(x)\phi(x)dx = (-1)^{|\alpha|} \int_{\mathcal{O}} u(x) D^\alpha \phi(x) dx,$$

For all test functions $\phi \in \mathcal{D}(\mathcal{O})$. Above $D^\alpha \phi = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} \phi$. Note that when the weak derivative $D^\alpha u$ exists, it is defined only up to a set of measure zero. So any point-wise statements to be made about $D^\alpha u$ is understood to only hold almost surely. Most of classical differential calculus can be reproduced for weak derivatives (e.g. the product rule, the chain rule).

Definition A.3. *The Sobolev space $H^s(\mathcal{O})$, $s \in \mathbb{N}$, is defined as the set of all functions $u \in L^2(\mathcal{O})$ with weak derivatives $D^\alpha u \in L^2(\mathcal{O})$ up to the order $|\alpha| \le s$.*

The above definition can be generalised for functions $u \in L^p(\mathcal{O})$, $1 \le p \le \infty$, and the resulting Sobolev spaces are usually denoted by $W^{s,p}(\mathcal{O})$. In this course we only consider $L^2(\mathcal{O})$ Sobolev spaces. The Sobolev spaces $H^s(\mathcal{O})$ are Banach spaces with the norm

$$\|u\|_{H^s} = \left(\sum_{|\alpha| \le s} \|D^\alpha u\|^2_{L^2(\mathcal{O})} dx \right)^{\frac{1}{2}} \tag{A.1}$$

The Sobolev spaces are separable Hilbert spaces with inner product

$$\langle u, v \rangle_{H^s} = \sum_{|\alpha| \le s} \langle D^\alpha u, D^\alpha v \rangle_{L^2} = \sum_{|\alpha| \le s} \int_{\mathcal{O}} D^\alpha u(x) D^\alpha v(x) dx,$$

for all $u, v \in H^s(\mathcal{O})$.

Definition A.4. *The spaces $H^s_0(\mathcal{O})$ are the closure of $C^\infty_0(\mathcal{O})$ under the Sobolev norm* (A.1)*.*

The spaces $H^s_0(\mathcal{O})$ is a closed subspace of $H^s(\mathcal{O})$. If $\mathcal{O} = \mathbb{R}^d$ then $H^s_0(\mathcal{O}) = H^s(\mathcal{O})$. We can define $H^1_0(\mathcal{O})$ also through Trace Theorem (see [5, Section 5.5]) which states that there is a continuous linear mapping tr : $H^1(\mathcal{O}) \to L^2(\partial\mathcal{O})$ called the trace operator. In this sense, we say that functions from $H^1(\mathcal{O})$ have traces (boundary values) in $L^2(\partial\mathcal{O})$ and

$$H^1_0(\mathcal{O}) = \{u \in H^1(\mathcal{O}) : u = 0 \text{ in } \partial\mathcal{O}\}.$$

As defined above, Sobolev spaces concern integer numbers of derivatives. However, the concept can be extended to fractional derivatives using Fourier transform.

Definition A.5. *Assume $0 \le s < \infty$ and $u \in L^2(\mathbb{R}^d)$. Then $u \in H^s(\mathbb{R}^d)$ if $(1 + |\xi|^s)\widehat{u} \in L^2(\mathbb{R}^d)$. The Sobolev norm is given by*

$$\|u\|_{H^s} = \|(1 + |\cdot|^s)\widehat{u}\|_{L^2},$$

where $\widehat{u} = \mathcal{F}(u)$ is the Fourier transform. Note that for a positive integer s, the above definition agrees with the definition given by the weak derivatives. For $s < 0$, we define $H^s(\mathbb{R}^d)$ via duality. The resulting spaces are separable for all $s \in \mathbb{R}$. If $\mathcal{O} \subset \mathbb{R}^d$ then $H^{-1}(\mathcal{O})$ is the dual space of $H_0^1(\mathcal{O})$.

In these notes we often consider $u \in L^2(\mathbb{T}^d)$, $\mathbb{T}^d$ being the d-dimensional unit torus, found by identifying opposite faces of the unit cube $[0,1]^d$. In this periodic case the Sobolev norm of the space $H(\mathbb{T}^d)$ can be written as

$$\|u\|_{H^s} = \sum_{\ell \in \mathbb{Z}^d} (1 + |\ell|^s)^2 \widehat{u}(\ell)^2.$$

We define the Laplace operator $\Delta = \nabla \cdot \nabla$ as $\Delta u = \sum_{i=1}^{d} \frac{\partial_i u}{\partial x_i^2}$ and note that the eigenvalues of $(I - \Delta)$ with domain $H^2(\mathbb{T}^d)$ are simply $1 + 4\pi^2|\ell|^2$, for $\ell \in \mathbb{Z}^d$. The fractional powers of $(I + \Delta)$ are defined as follows

$$(I - \Delta)^\gamma u = \sum_{\ell \in \mathbb{Z}^d} (1 + |\ell|^{2\gamma}) \widehat{u}(\ell) \phi_\ell,$$

where ϕ_k are the eigenvectors of $-\Delta$ in $\mathbb{T}^d$, that form the orthonormal basis of $L^2(\mathbb{T}^d)$. We see that on the torus $H^s = \mathcal{D}((I + \Delta)^{\frac{s}{2}})$ and we have $\|u\|_{H^s} = \|(I + \Delta)^{\frac{s}{2}} u\|_{L^2}$. We also note that that $(1 - \Delta)^{-r} : H^t(\mathbb{T}^d) \to H^{t+r}(\mathbb{T}^d)$ for all $t, r \in \mathbb{R}$.

References

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, 1975.

[2] V. I. Bogachev, *Gaussian measures*, vol. 62 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 1998.

[3] G. Da Prato and J. Zabczyk, *Stochastic equations in infinite dimensions*, vol. 152, Cambridge university press, 2014.

[4] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, Handbook of Uncertainty Quantification, (2016), pp. 1–118.

[5] L. C. Evans, *Partial differential equations*, American Mathematical Society, Providence, RI, 1998.

[6] M. Hairer, *An introduction to stochastic PDEs*, arXiv preprint arXiv:0907.4178, (2009).

[7] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.

[8] O. Kallenberg, *Foundations of modern probability theory*, Springer, 1997.

[9] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559.